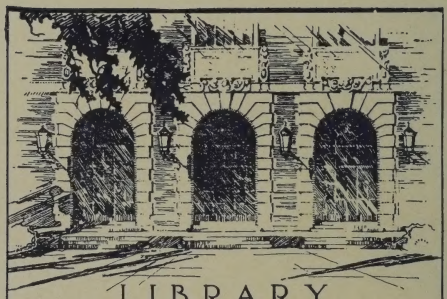
APR 27
1970

M



LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes*, all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

DIGITAL COMPUTER LABORATORY
LIBRARY

ALLERTON CONFERENCE ON CIRCUIT & SYSTEM THEORY: PROCEEDINGS

1969

1 WEEK CHECKOUT

Proceedings
SEVENTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 8-10, 1969

University of Illinois
Allerton House, Monticello, Illinois
Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois and the
Circuit Theory and Automatic Control Groups of the
Institute of Electrical & Electronics Engineers

IEEE Catalog No. 69 C 48-CT

#102, DCS Library

PROCEEDINGS
SEVENTH ANNUAL ALLERTON CONFERENCE
ON CIRCUIT AND SYSTEM THEORY

EDITED BY

A. H. HADDAD
G. METZE
T. N. TRICK
Conference Co-Chairmen

Conference held

October 8-10, 1969
Allerton House
Monticello, Illinois

Sponsored by the

DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
Urbana, Illinois

and the

CIRCUIT THEORY AND AUTOMATIC CONTROL GROUPS
of the
INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS

FOREWORD

The Seventh Annual Allerton Conference on Circuit and System Theory was held at Allerton House from October 8-10, 1969. These Proceedings are a permanent record of the papers presented at that Conference. The papers describe new results in information and coding theory, digital systems, communication sciences, control systems, network theory, and computer-aided design.

In addition to the technical sessions, the remotemess of Allerton Park continues to provide a setting fruitful to the exchange of ideas throughout the Conference.

The Conference was co-sponsored by the Department of Electrical Engineering and the Coordinated Science Laboratory of the University of Illinois, and the IEEE Groups on Circuit Theory and Automatic Control. We acknowledge the encouragement and support of Professor W. Dale Compton, Director of the Coordinated Science Laboratory, and Professor E. C. Jordan, Head of the Department of Electrical Engineering at the University of Illinois. Special thanks are due Rosa Townsend, Barbara Champagne, and Rose Lane who assisted with the Conference arrangements, and Mrs. Divona Keel who assisted with the editing of this Proceedings. Also, we thank Mr. Harold B. Lawler for his help with certain business aspects of the Conference. In addition, we are indebted to our colleagues for their vital assistance: R. T. Chien, J. T. Cordaro, Jr., Jose B. Cruz, Jr., Petar V. Kokotović, Wataru Mayeda, William R. Perkins, Franco Preparata, and Nelson Wax.

A. H. Haddad
G. Metze
T. N. Trick

TABLE OF CONTENTS

Foreword iii

A. GENERAL SESSION I
Chairman: Gernot Metze, University of Illinois

ON FEATURE ENCODING IN PICTURE PROCESSING BY COMPUTER . . 1
Julius T. Tou, University of Florida

B-1. CONTROL THEORY I
Chairman: A. J. Grammaticos, New York University

ON THE CONTROLLABILITY OF LINEAR SYSTEMS WITH DELAY IN CONTROL 19
Dong Hak Chyung, The University of Iowa

OBSERVABILITY: ITS ROLE IN CONTROL AND ESTIMATION THEORY 23
R. E. Klein, University of Illinois

A MODIFIED LIAPUNOV DESIGN FOR MODEL REFERENCE ADAPTIVE CONTROL SYSTEMS 32
J. W. Gilbart and R. V. Monopoli, University of Massachusetts

B-2. OPTIMAL CONTROL OF STOCHASTIC SYSTEMS
Chairman: M. K. Sain, University of Notre Dame

OPTIMAL CONTROL OF A CLASS OF STOCHASTIC SYSTEMS WITH UNKNOWN PARAMETERS 34
R. L. Kashyap, Purdue University

OPTIMAL CONTROL OF A LARGE CLASS OF STOCHASTIC DIFFERENTIAL SYSTEMS WITH FINITE MEMORY 44
N. U. Ahmed, University of Ottawa, Canada

AN OPTIMAL CONTROLLER FOR A PLANT WITH RANDOM PARAMETERS 54
H. Jon Thomas, Rochester Institute of Technology and John B. Lewis, Pennsylvania State University

B-3. DIGITAL FILTERS
Chairman: T. N. Trick, University of Illinois

A THEOREM CONCERNING LIMIT CYCLES IN DIGITAL FILTERS . . 63
I. W. Sandberg, Bell Telephone Laboratories, Inc.

AN ANALYSIS OF LIMIT CYCLES DUE TO MULTIPLICATION ROUNDING IN RECURSIVE DIGITAL (SUB)FILTERS 69
Leland B. Jackson, Bell Telephone Laboratories, Inc.

THE WEIGHTING MATRIX FORM OF THE DISCRETE OPTIMAL FILTER 79
Samuel L. Fagin, Samuel L. Fagin Associates, New York

B-4. IDENTIFICATION OF SYSTEMS
Chairman: C. A. Desoer, University of California, Berkeley

RESPONSE IDENTIFICATION OF DISTRIBUTED SYSTEMS WITH NOISY MEASUREMENTS AT FINITE POINTS 90
Paul C. Badavas, The Foxboro Co., Foxboro, Mass. and George N. Saridis, Purdue University

A RESULT ON THE IDENTIFICATION OF PARAMETERS IN DISTRIBUTED SYSTEMS REPRESENTED BY HYPERBOLIC EQUATIONS . . . 100
M. J. Wozny, Purdue University and W. T. Carpenter, Hughes Aircraft Co., Conoga Park, Calif.

ESTIMATION AND IDENTIFICATION OF NONLINEAR SYSTEMS WITH CONSTRAINTS 105
D. Isaacs and W. Lochrie, University of California, Irvine

B-5. CODING THEORY
Chairman: R. T. Chien, University of Illinois

MAJORITY LOGIC DECODING FOR LINEAR CODES 114
Paul Laity, Control Data Corp., St. Paul, Minnesota and K. C. Wang, Washington State University

A LARGE CLASS OF CYCLIC AND SHORTENED CYCLIC BINARY CODES FOR MULTIPLE ERROR CORRECTION 119
Harry T. Hsu, Pennsylvania State University

MAXIMUM LIKELIHOOD AND SEQUENTIAL DECODING OF SHORT CONSTRAINT LENGTH CONVOLUTIONAL CODES 130
F. Jelinek and L. R. Bahl, IBM

C-1. OPTIMAL CONTROL I
Chairman: Col. Bernard C. Morgan, Jr. Wright-Patterson Air Force Base

QUASI-OPTIMAL CONTROL WITH NONPERIODIC PIECEWISE-CONSTANT FEEDBACK GAINS 140
Louis R. Nardizzi, Southern Methodist University

A CONCISE EXPRESSION FOR THE SWITCHING COEFFICIENTS FOR POSITIVE-NEGATIVE FEEDBACK SYSTEMS 147
John Paul Hyde, USAF Academy

APPLICATION OF GALERKIN'S METHOD FOR THE DETERMINATION OF TIME-OPTIMAL CONTROL OF DISTRIBUTED-PARAMETER SYSTEMS . . . 156
S. S. Prabhu and I. McCausland, University of Toronto, Canada

OPTIMIZATION OF NONLINEAR CHARACTERISTICS 166
M. K. Inan and C. A. Desoer, University of California, Berkeley

A NEAR-OPTIMAL CONTROLLER FOR NONLINEAR SYSTEMS 169
Mohammad Jamshidi, University of Illinois

OPTIMAL SOLUTION OF TWO POINT BOUNDARY VALUE PROBLEMS WITH Q-TH ORDER STATE SPACE INEQUALITY CONSTRAINTS 181
Charles C. Schooler and Gerry L. Kelly, University of Kansas

A STUDY OF THE OUTPUT REGULATOR PROBLEM FOR LINEAR SYSTEMS WITH INPUT VECTOR 186
Umberto Di Caprio and Paul P. Wang, Duke University

C-2. COMPUTER-AIDED ANALYSIS
Chairman: Gabor C. Temes, University of California, Los Angeles

LINEAR PROGRAMMING APPLIED TO THE COMPUTER GENERATION OF EQUIVALENT CIRCUITS 188
I. D. Rapport, Case Western Reserve University and M. A. Murray-Lasso, Universidad Nacional Autonoma de Mexico

SYMBOLIC NETWORK FUNCTIONS BY A SINGLE PATH-FINDING ALGORITHM 196
P. M. Lin and G. E. Alderson, Purdue University

FAST COMPUTER ANALYSIS OF NONLINEAR SYSTEMS 206
R. H. Allen and A. F. Fath, Boeing Scientific Research Laboratories

COMPUTATIONAL STABILITY IN SYSTEM ANALYSIS 216
Bertil S. M. Granborg, University of Hawaii

IDENTIFICATION, CLASSIFICATION, AND REDUCTION OF COMPUTATIONAL ERRORS IN THE NASAP FLOWGRAPH METHOD 225
Dennis Sesar, Hughes Aircraft Co. and L. P. McNamee, University of California, Los Angeles

STATE SPACE ANALYSIS OF TIME-VARYING SYSTEM BY DIGITAL COMPUTER 235
Samuel C. Lee, New York University

C-3. DIGITAL SYSTEMS
Chairman: W. J. Poppelbaum, University of Illinois

DESIGN OF ASYNCHRONOUS SEQUENTIAL MACHINES WITH FAULT-DETECTION CAPABILITIES 237
K. C. Wang, Washington State University

RELIABILITY MODELING AND ANALYSIS OF A DYNAMIC TMR SYSTEM UTILIZING STANDBY SPARES 243
Francis Parkash Mathur, Jet Propulsion Laboratory

ELIMINATION OF SPURIOUS PULSES FOR TIME SLOT DECODERS . . 253
Jon G. Bredeson and Paul T. Hulina, Pennsylvania State University

THE USE OF CONNECTION GRAPHS FOR THE DETECTION OF DIGITAL FAULTS 261
Donald R. Schertz, Bradley University and Gernot A. Metze, University of Illinois

FAULT DIAGNOSIS OF COMBINATIONAL CELLULAR ARRAYS 272
Sharad Chandra Seth, University of Illinois

IMPLEMENTATION AND EXTENSION OF MULTI-DIMENSIONAL PATH SENSITIZING IN A SIMULATION AND DIAGNOSIS SYSTEM 284
S. A. Szygenda and G. C. Goldbogen, University of Missouri, Rolla

C-4. NONLINEAR SYSTEMS
Chairman: B. J. Leon, Purdue University

SEQUENTIAL ESTIMATION FOR DISCRETE-TIME NONLINEAR SYSTEMS 293
J. S. Meditch, Boeing Scientific Research Laboratories

A SECOND ORDER FILTER FOR STATE ESTIMATION IN NONLINEAR DISCRETE SYSTEMS 303
Chaw-Bing Chang and Hinrich R. Martens, State University of New York at Buffalo

STABILIZATION OF A CLASS OF NONLINEAR SYSTEMS USING A LINEAR SYSTEM ESTIMATOR 313
Robert E. Strane, Owens-Illinois, Inc., Pittsburgh, Pa. and William G. Vogt, University of Pittsburgh

AN APPROACH FOR FINDING THE SINUSOIDAL STEADY STATE RESPONSE OF NONLINEAR SYSTEMS 323
John C. Lindenlaub, Purdue University

A HYBRID APPROACH TO THE PROBLEM OF NONLINEAR FUNCTION GENERATION 330
R. N. Nilsen, University of California, Los Angeles

DIRECT MODELING OF NONLINEAR SYSTEMS USING HYBRID COMPUTER OPTIMIZATION TECHNIQUES 339
Charles H. Beck, Robert L. Drake and Ming H. Kuo, Tulane University

A PHASE-LOCKED LOOP WITH QUASI-LINEAR PHASE DETECTOR . . . 349
J. H. Yang, Page Communications Engineers, Washington, D. C. and S. S. Wolff, The Johns Hopkins University

C-5. SAMPLED DATA SYSTEMS
Chairman: B. C. Kuo, University of Illinois

CONSTANT DIRECTIONS OF THE RICCATI EQUATION 358
David Rappaport and Richard S. Bucy, University of Southern California, Los Angeles

ON THE CONTROLLABILITY OF DISCRETE-TIME-VARYING LINEAR SYSTEMS 368
Anthony J. Grammaticos, New York University

A SAMPLING METHOD FOR IDENTIFICATION OF LINEAR CONTINUOUS SYSTEMS 376
I. Eldem and G. J. Thaler, Naval Postgraduate School

DETERMINATION OF THE STATE TRANSITION MATRIX OF A SAMPLED-DATA CONTROL SYSTEM FROM MEASUREMENTS OF THE INPUT AND OUTPUT STATES 378
Willie L. McDaniel, Jr. and Jerrel R. Mitchell, Mississippi State University

A MAXIMUM PRINCIPLE FOR SAMPLED DATA SYSTEMS WITH TRAJECTORY CONSTRAINTS 380
Juergen Bruckner, Battelle Memorial Institute, Columbus, Ohio and Sherman Wu, Martin Marietta Corp., Orlando, Florida

ON QUANTIZED CONTROL SIGNAL SYSTEMS 382
R. Mark Havira and John B. Lewis, Pennsylvania State University

FIXED LAG SMOOTHING IN LINEAR DISCRETE SYSTEMS WITH TIME DELAYS 389
Roland Priemer, University of Illinois, Chicago and Andre G. Vacroux, Illinois Institute of Technology

D. JOINT MEETING with the Central Illinois Chapter of the IEEE Group on Circuit Theory and the Chicago Area Chapter of the IEEE Group on Information Theory
Chairman: T. N. Trick, University of Illinois

PROJECTS IN THE HARDWARE RESEARCH GROUP OF THE DEPARTMENT OF COMPUTER SCIENCE AT THE UNIVERSITY OF ILLINOIS 391
W. J. Poppelbaum, University of Illinois

E. GENERAL SESSION II
Chairman: J. B. Cruz, Jr., University of Illinois

A RESULT FOR LINEAR SYSTEMS DEFINED OVER ARBITRARY FIELDS: EQUIVALENCE OF CONTROLLABILITY AND POLE ASSIGNMENT 396
Sanjoy K. Mitter, Massachusetts Institute of Technology

NOVEL APPLICATIONS OF SCHWARZ'S INEQUALITY 399
A. Papoulis, Polytechnic Institute of Brooklyn

F-1. STOCHASTIC SYSTEMS
Chairman: Major John P. Hyde, USAF Academy

SOME NEW COMPUTATIONAL RESULTS FOR MINIMUM VARIANCE CONTROLLERS 408
Michael K. Sain and Stanley R. Liberty, University of Notre Dame

AN ANALYTICAL STUDY OF THE AGGREGATE TRANSFER FUNCTION FOR LINEAR STOCHASTIC SYSTEMS 418
Paul P. Wang, Duke University

THE OPTIMIZATION AND SIMULATION OF LINEAR STOCHASTIC SYSTEMS USING A STOCHASTIC MAXIMUM PRINCIPLE 428
Norman C. Fromm, Electronics Associates, Inc., Princeton, New Jersey and Earl D. Eyman, University of Iowa

A SUBOPTIMAL FILTER AND FILTER SENSITIVITY FOR STOCHASTIC PARAMETER SYSTEMS 440
Ronald L. Klein and Earl D. Eyman, University of Iowa

F-2. INFORMATION THEORY I
Chairman: J. T. Cordaro, Jr., University of Illinois

ON THE DETECTION OF NON-GAUSSIAN SIGNALS 450
J. J. Appel and E. N. Protonotarios, Columbia University

PERFORMANCE OF TWO SUBOPTIMUM DETECTORS AND SIGNAL SELECTION IN GAUSSIAN AND IMPULSIVE NOISE 459
Jack Kapp and Ludwik Kurz, New York University

SEQUENTIAL ESTIMATION FOR NONLINEAR SYSTEMS WITH GAUSS-MARKOV OBSERVATION NOISE 467
N. K. Loh, The University of Iowa

F-3. GRAPH THEORY
Chairman: W. Mayeda, University of Illinois

ON REALIZATION OF A CIRCUIT MATRIX 477
Shu-Gar Chan, Naval Postgraduate School, Monterey, California and James F. Nale, Naval Ships Engr. Center, Washington, D. C.

REARRANGEABILITY OF TRIANGULAR CONNECTION NETWORKS . . . 487
M. C. Paull, Rutgers University

ON CERTAIN OPTIMIZATION PROBLEMS IN COMMUNICATION NETWORKS 497
S. Y. Chan, University of California, Berkeley and San Diego State College and W. S. Chou, University of California, Berkeley and Network Analysis Corporation, Glen Cove, New York

FAULT DIAGNOSIS OF LINEAR SYSTEMS 503
M. F-S Lin and S. Park Chan, University of Santa Clara

F-4. IDENTIFICATION OF SYSTEMS AND SIGNALS
Chairman: J. C. Lindenlaub, Purdue University

STATE ESTIMATION, SYSTEM IDENTIFICATION, AND SIGNAL DETECTION 511
Kevin A. Clements, Singer-General Precision, Inc., Little Falls, New Jersey

A GENERAL APPROACH TO DISCRETE SYSTEM IDENTIFICATION . . 521
K. L. Suryanarayanan, A. C. Soudack and S. G. Rao, University of British Columbia

A NON-PARAMETRIC RECURSIVE IDENTIFICATION ALGORITHM FOR FINITE-MEMORY LINEAR SYSTEMS 523
Y. Bar-Shalom and S. C. Schwartz, Princeton University

F-5. PATTERN RECOGNITION
Chairman: K. S. Fu, Purdue University

UNSUPERVISED LEARNING STRUCTURE AND PARAMETER ADAPTIVE PATTERN RECOGNITION. PART I: INDEPENDENT SIGNALLING, CONTINUOUS DATA 533
D. G. Lainiotis, The University of Texas at Austin

AN ADAPTIVE SCHEME FOR PATTERN RECOGNITION 536
R. L. Lux and D. C. Lai, University of Vermont

STATISTICAL RECOGNIZER WITH ON-LINE FEATURE SELECTION AND ORDERING 538
Chi-hau Chen, Southeastern Massachusetts University

G-1. OPTIMAL CONTROL II
Chairman: J. B. Lewis, Pennsylvania State University

DESIGN OF SENSITIVITY AND STATE VARIABLE FEEDBACK OPTIMAL CONTROL SYSTEMS 539
Ronald E. Higginbotham, TRW Systems Group, Houston, Texas

ON THE SENSITIVITY OF THE LINEAR STATE REGULATOR 549
Dennis F. Wilkie, Ford Motor Company and Henk M. van Schieveen, University of Michigan

THE DESIGN OF AN OPTIMAL SYSTEM SUBJECT TO PARAMETER VARIATIONS 564
Fred J. Taylor, Texas Instruments, Dallas, Texas

AN ADAPTIVE APPROACH TO THE TRAJECTORY SENSITIVITY DESIGN OF OPTIMAL REGULATORS 574
Sira G. Rao and A. C. Soudack, University of British Columbia

SOME RESULTS ON THE USE OF REDUCED MODELS IN THE CONTROL OF LINEAR DYNAMICAL SYSTEMS 576
D. Mitra, Bell Telephone Laboratories, Inc.

CONJUGATE GRADIENT MINIMIZATION TECHNIQUE IN PROCESS OPTIMIZATION 586
S. Atluri, Canadian Westinghouse Company Limited

A PURSUIT-EVASION GAME WITH MULTIPLE PURSUERS 595
David M. Salmon, Wolf Management Services, Palo Alto, California

G-2. SYSTEM THEORY
Chairman: W. J. Perkins, University of Illinois

INPUT-OUTPUT PROPERTIES OF LINEAR DISCRETE SYSTEMS . . . 605
C. A. Desoer, University of California, Berkeley and M. Y. Wu, University of Colorado

INPUT-OUTPUT PROPERTIES OF NONLINEAR DISCRETE SYSTEMS . . 610
M. Y. Wu, University of Colorado and C. A. Desoer, University of California, Berkeley

SOME RESULTS IN PASSIVITY AND STABILITY OF SYSTEMS WITH A STATE REPRESENTATION 616
R. F. Estrada, Bell Telephone Laboratories, Inc. and C. A. Desoer, University of California, Berkeley

A CLASS OF SYSTEM ESTIMATORS FOR ROUTING PROBLEMS 626
Marlin H. Mickle, Ronald G. Hoelzeman and James T. Cain, University of Pittsburgh

A SYSTEM THEORY FRAMEWORK FOR MODELLING SECTORAL PRICES IN AN INTERREGIONAL ECONOMY 630
Prem S. Satsangi, Indian Institute of Technology and Jack B. Ellis, University of Waterloo

REDUCIBLE MATRICES AND THEIR APPLICATION TO COMPUTER ANALYSIS OF LINEAR TIME INVARIANT SYSTEMS 641
D. Syverson and J. Simes, Sacramento State College

A NEW APPROACH TO THE EVALUATION OF THE TRANSITION MATRIX 650
Ilia Kaufman, University of Waterloo, Canada

G-3. SWITCHING THEORY
Chairman: D. R. Schertz, Bradley University

ON A HOMOMORPHISM OF THE DE BRUIJN GRAPH AND ITS APPLICATIONS TO THE DESIGN OF FEEDBACK SHIFT REGISTERS. . 652
Abraham Lempel, Sperry Rand Research Center, on leave from the Technion-Israel Institute of Technology, Haifa, Israel

STATE ASSIGNMENTS FOR ASYNCHRONOUS SEQUENTIAL MACHINES. . 662
Chung-Jen Tan, IBM Watson Research Center

COST CONSIDERATIONS OF BASE N ADDERS USING N-VALUED LOGIC ELEMENTS 680
Anthony S. Wojcik and Gernot Metze, University of Illinois

A MARKOV MODEL FOR RELIABILITY OF PROBABILISTIC SEQUENTIAL CIRCUITS WITH MARKOV INPUTS 690
G. S. Glinski and Who-Kee Chung, University of Ottawa, Canada

ANALYSIS AND SYNTHESIS OF SEQUENTIAL FUZZY LOGIC CIRCUITS 692
Samuel C. Lee, New York University

G-4. STABILITY
Chairman: N. Wax, University of Illinois

$L_p(p \geq 1)$ AND ORLICZ STABILITY OF A LARGE CLASS OF NONLINEAR FEEDBACK SYSTEMS 702
N. U. Ahmed, University of Ottawa, Canada

STABILITY OF NONLINEAR RANDOMLY SAMPLED SYSTEMS 710
R. G. Agniel and E. I. Jury, University of California, Berkeley

THE ABSOLUTE STABILITY OF NONLINEAR SYSTEMS INVOLVING A TIME-VARYING GAIN 721
T. L. Chang and M. S. Davies, The University of British Columbia, Canada

ON THE STABILITY OF A CLASS OF CONTINUOUS STOCHASTIC SYSTEMS 730
J. E. Brandeberry, Wright State University and S. H. Wu, Martin Marietta Corp., Orlando, Florida

SET STABILITY OF DISCRETE DYNAMICAL SYSTEMS 732
James A. Heinen, Marquette University and Sherman H. Wu, Martin Marietta Corp., Orlando, Florida

ON THE STABILITY OF LURIE SYSTEMS WITH STOCHASTIC INPUT 734
S. Purkayastha, Technical Officer, Civil Aviation, New Delhi, India and A. K. Mahalanabis, Indian Institute of Technology, New Delhi, India

G-5. NETWORKS
Chairman: S. Park Chan, University of Santa Clara

ANALYSIS OF A NONLINEAR ACTIVE NETWORK WITH APPLICATIONS TO THE NEURISTOR. 742
N. J. Elias and M. S. Ghausi, New York University

GENERAL EXPRESSIONS FOR IDEAL TRANSMISSION FUNCTIONS, CHARACTERISTIC FUNCTIONS AND OPERATING VOLTAGE RATIOS FOR FILTER DESIGN USING TRANSFORMED VARIABLES 752
M. Sablatash, The University of Toronto, Canada

ANALYSIS AND SYNTHESIS OF FILTERS BY TRANSFORMATION OF MATRIX BINOMIALS 762
Thomas G. Marshall, Jr., Rutgers University

RICHARDS' THEORM FOR POSITIVE REAL MATRICES OF SEVERAL VARIABLES 773
Ahmed M. Soliman, University of Pittsburgh

REALIZATION OF A LOADED TWO-PORT WITH PRESCRIBED TRANSMISSION ZEROS AND DRIVING POINT FUNCTION 780
V. Ramachandran and M. N. S. Swamy, Sir George Williams University, Canada

SYNTHESIS OF TRANSFER FUNCTIONS USING DIFFERENTIAL INPUT AMPLIFIERS 782
K. Sankara Rao and V. G. K. Murti, Indian Institute of Technology, Madras, India

H. BANQUET TALK

Toastmaster: B. R. Myers, University of Notre Dame

Speaker: Victor J. Stone, Professor of Law, University of Illinois

Topic: Electronic Invasions of Privacy

J. GENERAL SESSION III
Chairman: M. E. Van Valkenburg, Princeton University

SIGNAL FILTERING WITH THE TRANSVERSAL EQUALIZER 792
R. W. Lucky, Bell Telephone Laboratories, Inc.

BANDLIMITED SYSTEMS WITH OPTIMUM TIME-RESPONSE 805
Gabor C. Temes, University of California, Los Angeles and Ampex

K-1. COMMUNICATION SYSTEMS AND INFORMATION RETRIEVAL
Chairman: R. T. Chien, University of Illinois

A MODULATION SCHEME FOR LARGE CARRIER SYSTEM TERMINALS. . 819
Sidney Darlington, Bell Telephone Laboratories, Inc.

SOME RESULTS ON INFORMATION SYSTEMS 826
David K. Chow, United Aircraft Research Laboratories

A NEW DEPARTURE IN INFORMATION PROCESSING BY FACSIMILE COMMUNICATION 835
Ira Kamen, Comfax Communications Corp. and Charles McDevitt, ComputerPix Corp.

K-2. CONTROL THEORY II
Chairman: P. Kokotović, University of Illinois

SENSITIVITY STUDIES OF ANALOG COMPUTER SIMULATIONS . . . 845
Hsueh-Hsiang Chi and Chi-Tsong Chen, State University of New York at Stony Brook

SINGULAR LINE COMPENSATORS 855
B. L. Shapiro and G. J. Thaler, Naval Post-graduate School, Monterey, California

A NEW CHART FOR ROOT LOCUS CONSTRUCTION OF GENERALIZED THIRD ORDER CONTROL SYSTEMS 865
M. Z. Mohammed and A. Y. Belal, University of Mosul, Iraq

K-3. INFORMATION THEORY II
Chairman: A. H. Haddad, University of Illinois

PERFORMANCE OF SEQUENTIAL DECODING WITH BIORTHOGONAL MODULATION AND Q-LEVEL QUANTIZATION 874
M. C. Kim and S. A. Tretter, University of Maryland

OPTIMAL OBSERVATION ALLOCATION IN ESTIMATION AND DETECTION 883
V. David Vandelinde, The Johns Hopkins University

IMPLEMENTATION OF THE WIENER MODEL 893
R. J. Kuhler and R. H. Seacat, Texas Technological University

M-ARY CLASSIFICATION -- THE CASE OF UNKNOWN A PRIORI INFORMATION 895
Yen-Bin Chen, IBM Systems Development Division

K-4. NASAP APPLICATIONS
Chairman: L. P. McNamee, University of California, Los Angeles

ON THE FORMULATION OF DICHOTOMOUS FLOWGRAPHS FOR NETWORKS HAVING EXCESS ELEMENTS 903
E. M. Rips, Newark College of Engineering

NASAP ANALYSIS OF NONLINEAR DC CIRCUITS 905
K. Haag, Illinois Institute of Technology

APPLICATION OF NASAP TO THE DESIGN OF MICROWAVE NETWORKS 915
G. J. Bronson and G. J. Herskowitz, Stevens Institute of Technology

ON FEATURE ENCODING IN PICTURE PROCESSING BY COMPUTER*

JULIUS T. TOU
University of Florida
Gainesville, Florida

Abstract - Representation of pictorial information via feature encoding is discussed. Several techniques for feature encoding in picture processing by computer are studied. These methods are applied to the classification of radiolaria, recognition of hand-written characters, and identification of fingerprints.

INTRODUCTION

In the study of information processing, one of the first problems we confront with is the precise representation of information. Experience tells us that a piece of information may be represented in various ways either in a sentence, or in a picture, or in a waveform, or in a map. The major means of characterizing a piece of information are (1) a language which consists of written or spoken symbols and a set of rules of grammar and semiotics, and (2) a picture which includes characters and maps. This paper is concerned with the representation of pictorial information via feature encoding. Several feature encoding techniques are discussed.

PICTURE PROCESSING BY COMPUTER

Picture processing is generally referred to as the operation which analyzes and transforms graphical and pictorial material, or as the operation which synthesizes and generates pictures from data or abstract rules alone, or as combination of these operations. The central problem in picture processing is the enormous amount of data involved.[1,2] In picture processing by computer, pictorial information is represented as a large two-dimensional array of numbers signifying light intensities, or as listings of characteristic attributes such as coordinates in a two- or three-dimensional space of significant points which define lines, contours, surfaces, and other relations, or as encoded strings of symbols characterizing the features of pictorial information.

Picture analysis and transformation consists of two major tasks: They are picture transmission, manipulation and reproduction, and pictorial recognition. In the former, pictures are digitized after transformation into analog electrical signals by a scanner. The digitized pictures are

*This work was supported in part by the Office of Naval Research under Contract N 00014-68-A0173-0001, NR 049-172, and in part by the National Science Foundation under grant GK-2786.

processed by the computer to achieve bandwidth compression in transmission, and the processed pictorial information is then enhanced and redisplayed as pictures at a remote terminal. Bandwidth is a precious commodity. The removal of redundant or irrelevant information from pictorial data is important in many communication systems. Redundant data can in principle be removed without loss of information. Pictorial recognition is concerned with the interpretation, identification and classification of the contents in a picture by analysis of the graphical material. Computer serves as a powerful tool for performing the analysis. Among the notable examples of pictorial recognition are the classification of fingerprints, recognition of hand-printed characters, and identification of targets in aerial photographs. While bandwidth compression emphasizes communication aspects, pictorial recognition emphasizes decision aspects.

The operation of picture synthesis and generation may be sub-divided into two types: (1) A picture is synthesized from data or physical specifications; (2) a picture is constructed from abstract ideas and specifications. Applications of data-to-picture processing may be found in topography and geography. A large collection of data is stored in the computer to form the data base. From the organized and stored information, maps and other pictorial displays showing population density, geographical terrain, economic levels, housing conditions, or resources distribution may be generated upon request by employing automatic information retrieval schemes. The abstraction-to-picture operation deals with the creation of graphs and pictures from literal statements, mathematical symbols, or abstract concepts. Computer may be used as a tool to construct line drawings and pictures from ideas, symbols and statements. This picture processing operation finds useful applications in industrial design, printed-circuits production, computer art, among other things.

In picture processing by computer, the first problem we shall encounter is the representation of pictorial information so as to enable the computer to understand, to acquire and to process it. Computer cannot process pictorial material directly, unless the pictorial information is converted into machine readable form. Before efficient processing takes place, the pictorial information is first precisely represented in terms of its features which are coded into a string of symbols. Such representation of pictorial information is referred to as feature encoding. The coded feature strings can be fed into the computer memory for processing. Computer files contain visual image resembling a half-tone picture. The operation which

may be performed within the computer includes smoothing, boundary detection, edge enhancement, object isolation, scene abstraction, object identification, and data reduction. These operations can be performed either locally or globally, and can be made conditional on the results of previous operations. According to the structure, pictures may be classified into two types: (1) Photographic pictures, and (2) graphic pictures or line drawings. Photographic pictures which are characterized by continuous grey levels contain poorly defined boundaries and extremely high information density. Graphic pictures which are characterized by binary level of intensity contain well-defined contours and low information density. This paper is concerned with feature encoding for graphic pictures.

FEATURE ENCODING

It has been recognized that the central problem in picture processing is the enormous amount of data involved. To process pictorial material by computer often requires a huge memory capacity. This problem may be alleviated via encoding. Pictorial information may be expressed in machine readable form by feature encoding. In this section we shall discuss five feature encoding techniques which may be used to advantage in picture processing and classification by computer. These feature encoding techniques are (1) boundary-segmentation approach, (2) rectangular-grid quantization, (3) radial-grid encoding, (4) syntactical-analysis approach, and (5) topological-string encoding. The main purposes of these techniques are to encode pictorial information into machine readable form and to reduce machine storage requirement by making use of the features in the pictorial material. Of the five feature encoding techniques some are useful in picture manipulation and reproduction, and the others find applications in picture analysis and pictorial recognition.

Boundary-segmentation Approach

The contours of graphic pictures or line drawings may be considered as consisting of a number of segments. Each segment is characterized by (a) the coordinates of the initial point, (b) the tangent vector, and (c) the curvature of the segment. When the length of a segment is chosen sufficiently small, the curved segment may be approximated by its chord, and the angle θ between the tangent vector and the chord becomes an approximate measure of the curvature of the segment. When the curvature-threshold for the segments is specified to maintain certain accuracy, the terminal point of each segment can readily be determined. The tangent vector is a

measure of the direction of the segment, and the curvature-threshold specifies the segment length which is long for segment with small curvature and is short for segment with large curvature. Each boundary segment is analyzed successively and is represented by a boundary characterization list.

Let the coordinates of the initial point of a graphic picture be (x_o, y_o), the tangent vector be T_1, and the chord vector be C_1 with components C_{1x} and C_{1y}. Then the first segment is represented by

$$< x_o, y_o, T_1, C_1 >.$$

The coordinates of the initial point of the second segment is given by $(x_o + C_{1x}, y_o + C_{1y})$. The tangent vector at this point and the chord vector for the second segment are determined and represented by T_2 and C_2, respectively. The second segment can be represented by

$$< x_o + C_{1x}, y_o + C_{1y}, T_2, C_2 >.$$

With the components of C_2 be denoted by C_{2x} and C_{2y}, the coordinates of the initial point of the third segment is given by

$$(x_o + C_{1x} + C_{2x}, y_o + C_{1y} + C_{2y}).$$

Let the tangent vector at this point and the chord vector for the third segment be T_3 and C_3, respectively. Then the third segment is represented by

$$< x_o + C_{1x} + C_{2x}, y_o + C_{1y} + C_{2y}, T_3, C_3 >.$$

Following the above analysis, the nth segment is represented by

$$< x_o + \sum_{i=1}^{n-1} C_{ix}, y_o + \sum_{i=1}^{n-1} C_{iy}, T_n, C_n >.$$

Rectangular-grid Quantization

In dealing with pictorial pattern recognition problems, such as fingerprint identification, often a need arises to describe an arbitrary curve or contour. The selection of the method of representation is based upon the considerations of accuracy in description, data economy, simplicity of utilization. Accuracy in description may be insured by the use of a few complex mathematical approximations to the curve or by the use of a number of simple approximating segments, as in the boundary-segmentation approach. Data economy may be achieved by an encoding process which makes best use of known ensemble information. Such encoding may assume different schemes for various processing task if maximum data economy is to be achieved. However, for ease in utilization a standard and simple encoding system should be used. The rectangular-grid quantization approach can reasonably satisfy the requirements for accuracy in description and simplicity of utilization. This feature encoding method involves two major steps: (1) Superimpose a rectangular grid on the curve, and (2) label as a curve point the grid node which is closest to each intersection of the grid and the curve. It is noted that the error due to this quantization is one half of the grid step. The use of the rectangular grid results in two standard distances between successive curve points, which are in the ratio of $\sqrt{2}$ to each other. One is the side and the other is the diagonal of a small square in the grid.

Instead of rectangular grid, the hexagonal grid can also be used to achieve quantization. The use of the hexagonal grid results in constant distance between successive curve points, but the angle is quantized into six levels instead of eight as in the rectangular grid case. The hexagonal grid system is to be preferred only if constant distance between successive curve points is an important factor. The use of the rectangular grid system results in finer angle quantization and is compatible with common machine scanning approaches.

Feature encoding by the grid quantization method results in spatial quantization noise which can vary as the grid location drifts with respect to the input curve from quantization to quantization. The grid-curve intersections which are approximately equidistant from the two nearest grid nodes are particularly sensitive to grid movement. Spatial distortion due to such noise may be reduced as the grid size is made small.

Radial-grid Encoding

To encode the continuous portion of a curve, one may make use of the fact that the points describing a continuous curve are not independent and

that successive points are adjacent to each other. We may design an efficient encoding scheme by exploiting the a priori information that each curve point must have one of its immediate grid neighbors as a next curve point. In the rectangular coordinates system, a curve point possesses eight nearest neighbors located at the terminals of a radial grid as shown in Figure 1. The next curve point must be one of the eight nearest neighbors. Thus the location of each succeeding curve point can be made dependent upon the preceding point by encoding only the direction data to each new point. Absolute coordinate data can be given for the first curve point of each continuous curve segment, if it is desired to fix the absolute location of the segment. An illustration of radial-grid encoding is shown in Figure 2. By ignoring the initial coordinates, the graphic picture is encoded as 170654. Straight-line approximations are used between curve points. The segment between two successive curve points is of length $a(2)^{i/2}$ where $i = 0$ for an even direction number, $i = 1$ for an odd direction number, and 'a' is a scale factor which determines the accuracy for each line segment approximation. Such an encoding scheme is simple and efficient. However, it is susceptible to error build-up during encoding.

The above feature encoding scheme facilitates manipulation of the curve by a digital computer.[3] Simple manipulations of graphic pictures such as rotation, enlarging and reduction can readily be performed. Rotation by a multiple N of 45° can be accomplished simply by replacing each chain element k by the chain element represented by the sum (k+N) modulo 8. When N is even, no distortion of a rectangular quantized curve results. When N is odd, there will be a distortion due to interchange of the element lengths a and $a\sqrt{2}$. Rotation by angles not a multiple of 45° is very difficult to perform if severe distortion is to be avoided. The enlarging operation can be performed symbolically by replacing each of the digits of coded representation by a set of identical digits. Thus, scale enlargement by a factor n of the coded curve can be accomplished by replacing each chain element by n identical elements. Overall reduction is most easily performed by changing the value of the scale factor 'a'. Selective reduction or contraction implies dropping chain elements out of the coded curve. This will usually result in distortion.

Operations in graphical picture tracing such as reverse tracing, closure testing, length and area finding, and intersection determination can be readily performed. Reverse tracing can be achieved by reversing each element direction in the chain and then reversing the sequence order within

the chain. For example, the reverse tracing of 170654 is 012435. Closure is determined by the chain residue. If a curve is closed then its residue is zero; if it is open the residue is not null. The approximate length of a curve is given by the number of even directed elements plus the square root of two times the number of odd directed elements. The area enclosed by the curve may be determined by the trapzoidal integration rule, with the area under a line segment directed to the right designated as positive and the area under a line segment directed to the left designated as negative. A nodal intersection can be determined by constructing line-segment sequences for the curves. The nodal intersection is the point common to both sequences. A non-nodal intersection may be determined in the same manner if the quantization level is doubled. By reducing the scale factor to one half, a non-nodal intersection becomes a nodal intersection.

Syntactical Analysis Approach

The boundary of a pictorial object, such as a chromosome and the skeletal remains of radiolaria found in marine sediment, may be considered as consisting of a set of characteristic segments. These segments are, for example, clockwise curves, counterclockwise curves, line segments, notches, clockwise bends, counterclockwise bends, caret, fin. By combining these segments in proper sequence, we may reproduce the approximate contour which retains the same general configuration and topological properties, although we cannot duplicate the original boundary. The synthesized contour may be expressed in terms of a list of symbols, each representing a characteristic segment. By making use of the syntactical definitions, we may describe the build-up parts of the pictorial object from combinations of these characteristic segments.

As an illustratration, we apply this approach to automatic recognition of the outline shape of organisms in oceanographic research. The contours of typical radiolaria may be characterized by eight basic segments: Clockwise curve, counterclockwise curve, line segment, notch, clockwise bend, counterclockwise bend, caret, and fin. Let these characteristic segments be denoted by A, B, C, D, E, F, G, and H, respectively, as shown in Figure 3a. The detailed discussion of syntactical analysis approach to automatic recognition and classification of radiolaria will be presented in a future paper. Here in view of space limitations we will only cover the basic notions. To facilitate illustration, we will restrict the syntactic analysis of type (b), type (d), type (f), and type (g) radiolaria, which are depicted in Figure 4.

From the basic segments we may derive several secondary segments which are characterized by syntactical recursive definitions. The derived secondary segments for types (b), (d), (f) and (g) radiolaria are bell, cup, and spike, as shown in Figure 3b. For the characterization of some other types of radiolaria, we need more derived secondary segments. The syntactical definition

$$< \text{bell} > :: = C^* < \text{bell} > \mid < \text{bell} > C^* \mid A$$

implies that the generic concept bell is defined as being a concatenated C type segment followed by a bell, or a bell followed by a concatenated C type segment, or an A type segment. In the above expression, the angular brackets < > imply that a generic name is enclosed, the symbol ":: =" means "is defined as being", and the symbol "|" denotes the "or" operation. The definition is recursive in that it is used repeatedly. We first identify the A type segment, and then build up the bell by repeated application of the definition to include the C type segments on the left and the C type segments on the right. The other two derived secondary segments are

$$< \text{cup} > :: = C^* < \text{cup} > \mid < \text{cup} > C^* \mid E \mid F$$

and

$$< \text{spike} > :: = C^* < \text{spike} > \mid < \text{spike} > C^* \mid G.$$

Furthermore, we may derive the syntax for the build-up parts. For types (b), (d), and (f) radiolaria, we have

$$< \text{cap} > :: = ED < \text{spike} > \mid < \text{spike} > DE$$

$$< \text{tail} > :: = < \text{spike} > B < \text{spike} > B < \text{spike} >$$

$$< \text{tower} > :: = DE < \text{cap} > \mid < \text{cap} > ED$$

$$< \text{claw} > :: = < \text{cup}_2 > < \text{bell}_1 > < \text{cup}_2 > < \text{bell}_1 > < \text{cup}_2 >$$

Then the types (b), (d), and (f) radiolaria can be expressed syntactically as

$$\text{Type (b) Rad} :: = < \text{cap} > < \text{cap} >$$

$$\text{Type (d) Rad} :: = < \text{bell}_1 > < \text{claw} >$$

$$\text{Type (f) Rad} :: = < \text{tower} > < \text{tail} >$$

The identification of a particular type of radiolaria may be carried out by means of a simple automaton. It is noted that, using the above feature encoding technique, type (b) and type (g) radiolaria are represented by

GC*DEDC*-2 and GC*DABEC*EC*EBADC*, respectively. An automaton to identify these two types of radiolaria is given in Figure 5. Application of syntactical analysis approach to chromosome classification are discussed in reference 4.

Topological-string Encoding

The features of certain pictorial material may be encoded in terms of a string of topological properties. Such feature encoding scheme has found applications in recognition of hand-written characters and in automatic fingerprint classification studies.

(a) Recognition of hand-written characters - In the character recognition work, a hand-written character may be represented by a measurement vector of eight dimension or a string of eight symbols. Each dimension or symbol describes the topological property of the character in an octant. The eight octants are generated by a radial grid with its center located near the center of the character as shown in Figure 6.

When no stroke appears in an octant, it is represented by the symbol a_o. Any stroke which crosses octants 3, 4, 7, 8 and can be approximated by a vertical stroke is represented by the symbol a_1. Any stroke which crosses octants 1, 2, 5, 6 and can be approximated by a horizontal stroke is represented by the symbol a_2. A stroke which terminates in an octant in the clockwise direction is represented by the symbol a_3; and a stroke which terminates in an octant in the counterclockwise direction is represented by the symbol a_4. A loop in an octant which opens at the right is represented by the symbol a_5; and a loop in an octant which opens at the left is represented by the symbol a_6.

Via a topological transformation, a hand-written character may be converted into a stylized configuration which retains the same topological properties. The elements of the coded string are the topological features of the character. As an illustration, consider the transformation exemplified in Figure 7.

The code string for this character is

$$< a_4,a_2 > < a_2,a_2 > < a_6 > < a_4 > < a_2 > < a_2 > < a_1 > < a_1 >$$

which starts from the first octant and follows a clockwise direction.

To facilitate computation, each of the above code symbols is assigned an appropriate number. We then simplify the classification operation by combining the code strings which belong to the same class to form one or

several sets of prototype feature vectors via clustering and training. The occurrence of the basic features carries different weights in different octants when the prototype feature vector is being formed. This feature extraction process results in one (sometimes several) optimum number for the topological features in each octant. Let this feature number be denoted by $F_i(j,k,\ell,m)$, where i = 1,2, ..., 8, denoting the ith octant, j,k,ℓ are the values assigned to the feature strokes in octant i, and m = 1 when there is no crossing of strokes in octant i, m = 2 when there is crossing of strokes in octant i. A computer program has been written to perform this transformation.

Consider the character shown in Figure 8. The code string is found to be

$$< 3 > < 8 > < 1 > < 11 > < 3,3 > < 3,9 > < 2 > < 2 >.$$

The computer program generates the feature numbers as

$F_1(3,1,1,1) = 1$ $\qquad$ $F_5(3,3,1,1) = 12$

$F_2(8,1,1,1) = 4$ $\qquad$ $F_6(3,9,1,1) = 9$

$F_3(1,1,1,1) = 7$ $\qquad$ $F_7(2,1,1,1) = 1$

$F_4(1,1,11,1) = 12$ $\qquad$ $F_8(2,1,1,1) = 1$

Hence the prototype feature vector for this class is

$$\underline{X}_6 = \text{col } (1,4,7,12,12,9,1,1)$$

A complete discussion of this approach to recognition of hand-written characters will be presented in a forthcoming paper.[5]

(b) Automatic fingerprint processing - In the automatic fingerprint interpretation system proposed by Tou and Hankley, the topological features of a fingerprint are encoded in terms of a string of symbols to form a topological code sentence. This coded information is stored in a data file for retrieval and identification purposes. The system converts input fingerprint into a digitized fingerprint matrix which is then coded into a line pattern. The topological coding subsystem "reads" the line pattern and generates a code sentence which describes the topological structure of the fingerprint. Ordering of code sentence in the data file is the basis for fingerprint classification.

The curves in the line pattern are directed downward from locally uppermost points which are labeled T_i and are called top nodes. These nodes are points of locally maximum curvature. The other types of nodes are line

intersection nodes, line end nodes, and radial nodes. A typical line pattern with marked nodes is illustrated in Figure 9.

The code sentences which are generated by the topological coding program are defined over an alphabet [T, L, R, B, 0, 1], where T, L, R, B, 0 denote top, left, right, bottom, and null, respectively, and 1 is used as a multiple postscript. The code sentence of a line pattern is composed of n words, where n is the number of completely coded line subgraphs emanating downward from the selected top nodes. Each word represents the topological structure of the connected subgraph read downward only from the corresponding top node. Word structure is defined recursively by the node syntax which is of the form

$$G_i ::= N_i \mid N_i(0) \mid N_i(G_1 \ldots G_f)$$

where N_i denotes the ith node. G_i is either the node N_i or a string $N_i(\ldots)$ which contains either the null symbol or else a list $G_1 \ldots G_f$ corresponding to the f nodes connected directly below the ith node ordered left to right. When a non-bottom node has been previously labeled, then that label is followed by a postscript 1 so as to identify the recurrence of the same label. Figure 10 illustrates some topological code words which follow the above node syntax. Several simple code sentences are derived in Figure 11. The topological code sentence for a sample fingerprint is given in Figure 12. The design of this automatic fingerprint interpretation system is discussed in reference 6.

CONCLUSION

Picture processing is generally referred to as the operation of analyzing and transforming graphical and pictorial material, or as the operation of synthesizing and generating pictures from data or abstract rules alone, or as combination of these operations. Picture analysis and transformation are concerned with the transmission, manipulation and reproduction of pictures and the recognition of pictorial patterns. A picture may be synthesized from data or physical specifications, and can be generated from abstract ideas. Computers may be used to perform spatial filtering, boundary detection, edge enhancement, object isolation, scene abstraction, object identification, and data reduction. In picture processing by computer, the pictorial information is first represented in terms of its features which are coded into a string of symbols. Five feature encoding techniques for graphic pictures are examined. Applications to radiolaria

classification, hand-written character recognition, and fingerprint interpretation are discussed. Linguistic approach is becoming a major theme in picture processing by computer. The development of a powerful programming language for picture processing seems desirable and necessary.

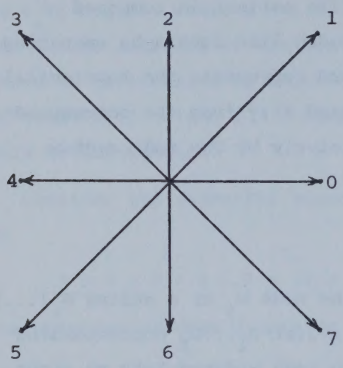


Figure 1 - Radial grid

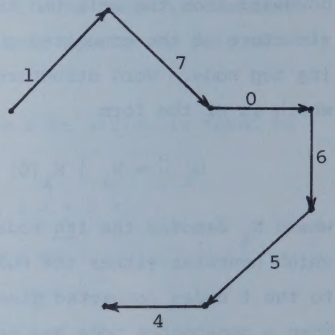


Figure 2 - Radial-grid encoding

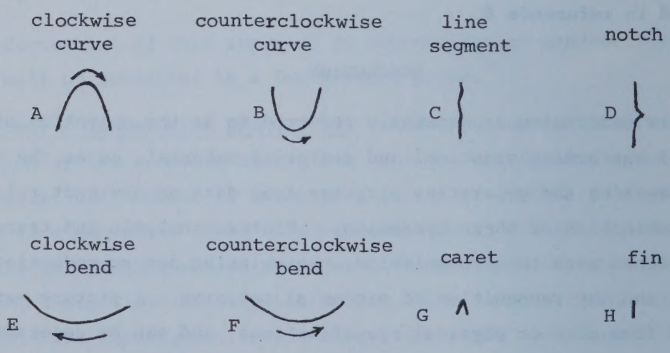


Figure 3a - Characteristic segments

REFERENCES

1. Hawkins, Joseph K., "Image Processing Principles and Techniques", in Advances in Information Systems Science, Vol. 3, Julius T. Tou (Ed.), Plenum Publishing Corporation, New York, 1970.

2. Harmon, Leon D. and Knowlton, Kenneth C., "Picture Processing by Computer", Science, Vol. 163, April, 1969.

3. Freeman, Herbert, "On the Encoding of Arbitrary Geometric Configuration", IRE Transactions on Electronic Computers, June, 1961.

4. Ledley, R. S., et al, "FIDAC: Film Input to Digital Automatic Computer and Associated Syntax-directed Pattern-recognition Programming System", in Optical and Electro-optical Information Processing, J. Tippett, et al (Eds.), MIT Press, 1965, pp. 591-614.

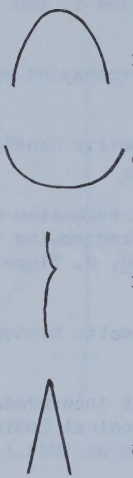
5. Tou, J. T. and Gonzalez, R., "A New Approach to Automatic Recognition of Hand-written Characters", to appear.

6. Hankley, W. J. and Tou, J. T., "Automatic Fingerprint Interpretation and Classification via Contextual Analysis and Topological Coding", in Pictorial Pattern Recognition, George C. Cheng, et al (Eds.), Thompson Book Company, Washington, D. C., 1968.

7. Tou, J. T., "Engineering Principles of Pattern Recognition", in Advances in Information Systems Science, Vol. 1, Julius T. Tou (Ed.), Plenum Publishing Corporation, New York, 1969.

8. Pankhurst, R., "Computer Graphics", in Advances in Information Systems Science, Vol. 3, Julius T. Tou (Ed.), Plenum Publishing Corporation, New York, 1970.

9. Rosenfeld, A., "Picture Processing by Computer", Computing Surveys, Vol. 1, No. 3, Sept. 1969.



Bell

$< bell_1 > :: = C^* < bell_1 > | < bell_1 > C^* |A$

$< bell_2 > :: = C^* < bell_2 > | < bell_2 > C^* |B$

Cup

$< cup_1 > :: = C^* < cup_1 > | < cup_2 > C^* |E$

$< cup_2 > :: = C^* < cup_2 > | < cup_2 > C^* |F$

Dent

$< dent > :: = C^* < dent > | < dent > C|CD|DC|D$

Spike

$< spike > :: = C^* < spike > | < spike > C^* |G$

Figure 3b. Derived secondary segments

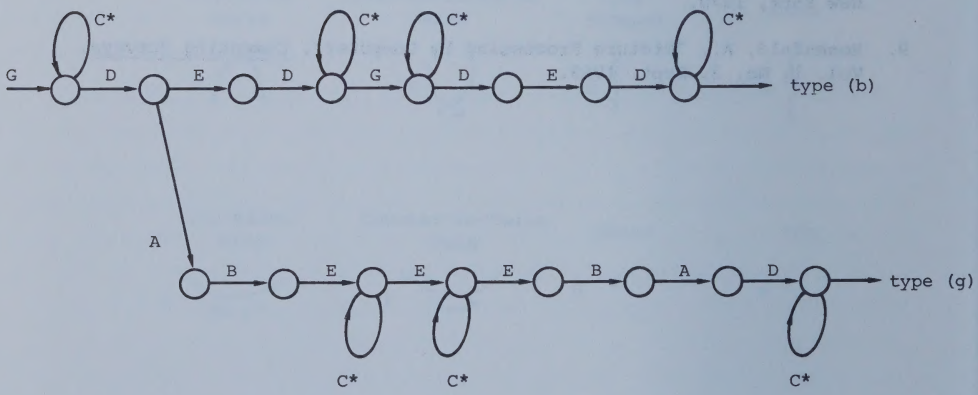


Figure 5 - Classification automaton

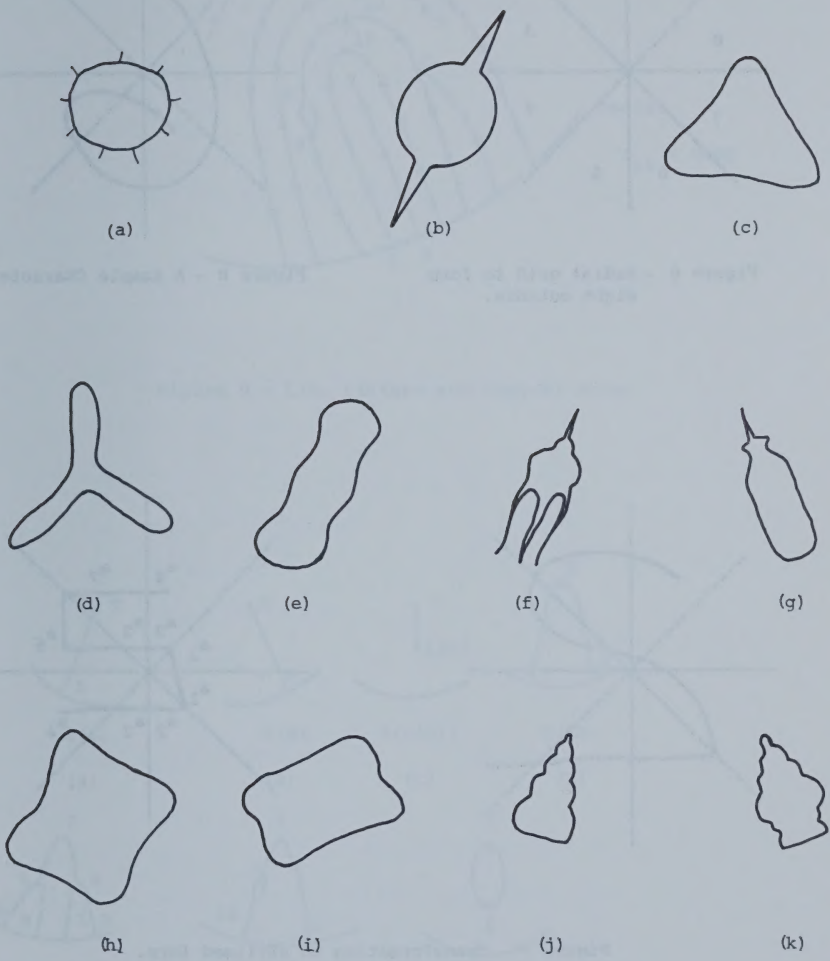


Figure 4 - Typical radiolaria

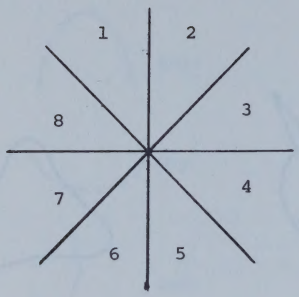


Figure 6 - Radial grid to form eight octants.

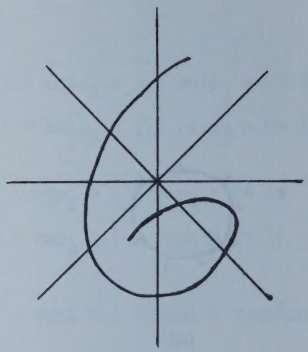
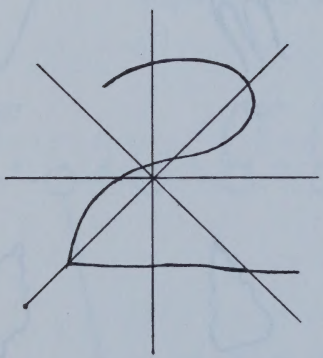


Figure 8 - A Sample Character.



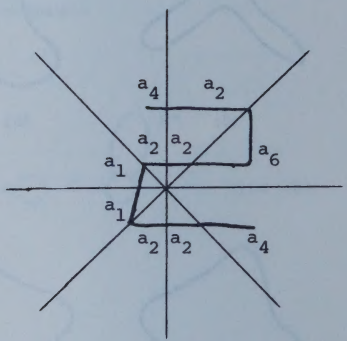


Figure 7 - Transformation to stylized form.

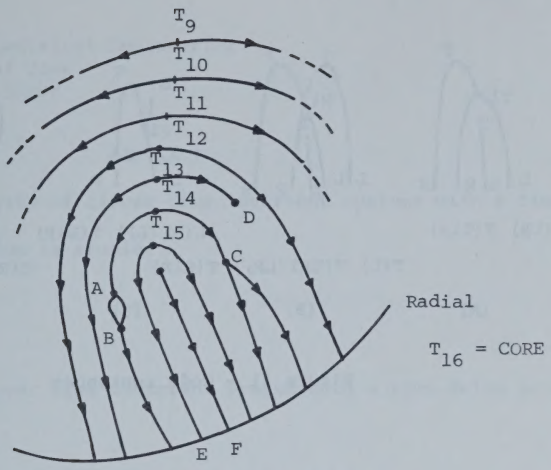


Figure 9 - Line pattern and typical nodes

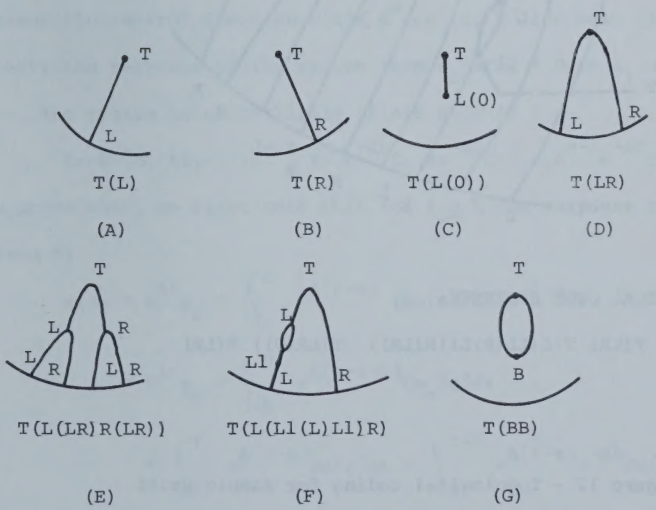
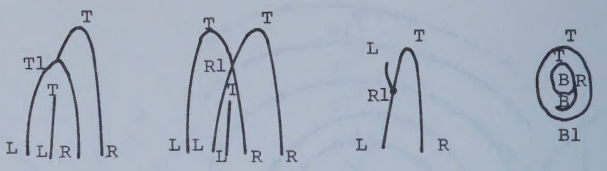


Figure 10 - Topological code words

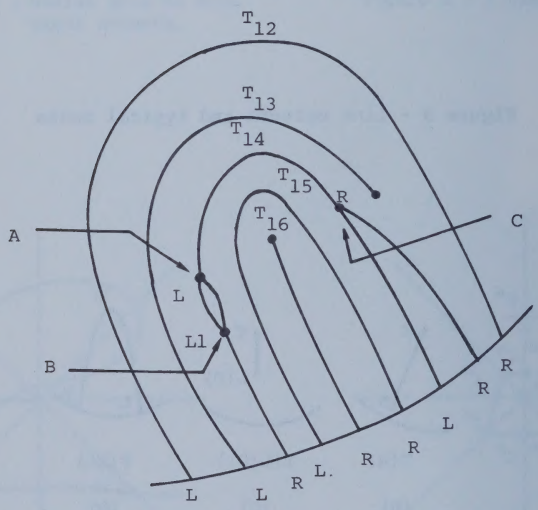


T(L) T1(LR) T(T1R) L(R1(L)) T(R1R)

T(L) T(LR1(LR)) T(R1R) T(RR(BB)) T(B1B1)

(A) (B) (C) (D)

Figure 11 - Code sentences



TOPOLOGICAL CODE SENTENCE:

T(R) T(LR) T(L(L1(R)L1)R(LR)) T(LR(0)) T(LR)

Figure 12 - Topological coding for sample print

ON THE CONTROLLABILITY OF LINEAR SYSTEMS WITH DELAY IN CONTROL

DONG HAK CHYUNG
Department of Electrical Engineering
The University of Iowa
Iowa City, Iowa 52240

ABSTRACT

The controllability of linear time invariant systems with a time delay in the control function is studied.

Consider the linear time-invariant system with a time delay in the control function

$$\dot{x}(t) = Ax(t)+Bu(t)+Cu(t-h)$$

where $x(t)$ is the n-dimensional vector state variable, $u(t)$ is the m-dimensional vector control function, $h > 0$ is the time delay, and A, B and C are constant matrices of compatible dimensions.

The system is called controllable if, for any given pair of points x_o, x_1 and initial control function $u_o(t)$ on $[-h, 0]$, there exists a bounded measurable control function $u(t) \subset R^m$ on $[0, T]$ for some finite time T which steers the response of the system from x_o at $t = 0$ to x_1 at $t = T$.

The system is controllable if and only if

$$\text{Rank } [B, AB, \ldots, A^{n-1}B, e^{-Ah}C, Ae^{-Ah}C, \ldots, A^{n-1}e^{-Ah}C] = n$$

To prove this, we first note that for $t \geq h$ the response of the system is given by

$$x(t) = e^{At}x_o + \int_o^t e^{A(t-s)}[Bu(s)+Cu(s-h)]ds$$

$$= e^{At}x_o + \int_{-h}^{o} e^{A(t-s-h)}Cu_o(s)ds$$

$$+ \int_o^t e^{A(t-s)}Bu(s)ds + \int_o^{t-h} e^{A(t-s)}e^{-Ah}Cu(s)ds$$

$$= e^{At}\left\{x_o + \int_{-h}^{o} e^{-A(s+h)}Cu_o(s)ds\right\}$$

$$+ \int_{o}^{t-h} e^{A(t-s)}[B+e^{-Ah}C]\,u(s)ds + \int_{t-h}^{t} e^{A(t-s)}Bu(s)ds$$

Now to show the necessity, suppose Rank $[B, AB,\ldots,A^{n-1}B, e^{-Ah}C, Ae^{-Ah}C,\ldots,A^{n-1}e^{-Ah}C] < n$. Then there exists a nonzero vector $v \in R^n$ such that $vB = vAB = \ldots = vA^{n-1}B = ve^{-Ah}C = vAe^{-Ah}C = \ldots = vA^{n-1}e^{-Ah}C = 0$. So by the Cayley-Hamilton theorem $vA^kB = 0$, $vA^ke^{-Ah}C = 0$ for all $k \geq 0$. This implies that $ve^{At}B = 0$, $ve^{At}e^{-Ah}C = 0$ for all real t.

Consider the initial point, for a given $u_o(s)$ on $[-h, 0]$,

$$x_o = -\int_{-h}^{o} e^{-A(s+h)}Cu_o(s)ds$$

Then

$$x(t) = \int_{o}^{t} e^{A(t-s)}Bu(s)ds + \int_{o}^{t-h} e^{A(t-s)}e^{-Ah}Cu(s)ds$$

and

$$vx(t) = \int_{o}^{t} ve^{A(t-s)}Bu(s)ds + \int_{o}^{t-h} ve^{A(t-s)}e^{-Ah}Cu(s)ds = 0$$

for all u(s) on $[0, t]$ for all t. Since $vx(t) = 0$ implies $v \neq x(t)$, if we let $x_1 = v$ then it is impossible to steer the response from the given x_o to the given $x_1 = v$, and so the system is not controllable. Hence if the system is controllable, then Rank $[B, AB,\ldots, A^{n-1}e^{-Ah}C] = n$.

Conversely, assume Rank $[B, AB,\ldots, A^{n-1}e^{-Ah}C] = n$ and suppose the system is not controllable. Then, for any given initial function $u_o(s)$ on $[-h, 0]$ there exists a pair of points x_o and x_1 such that no control can steer the response from x_o at $t = 0$ to x_1 at $t = 2h$, that is,

$$x_1 \neq x(2h) = e^{A(2h)}x_o + \int_{-h}^{o} e^{A(h-s)}Cu_o(s)ds$$

$$+ \int_{o}^{h} e^{A(h-s)}[B + e^{-Ah}C]\,u(s)ds + \int_{h}^{2h} e^{A(2h-s)}Bu(s)ds$$

and

$$w = x_1 - e^{A(2h)}x_o - \int_{-h}^{o} e^{A(h-s)}Cu_o(s)ds \neq 0$$

for all bounded measurable control function $u(s) \subset R^m$ on $[0, 2h]$.

Rearranging the equation above,

$$\int_o^h e^{A(2h-s)}[B + e^{-Ah}C]\, u(s)ds + \int_h^{2h} e^{A(2h-s)}Bu(s)ds$$

$$\neq w = x_1 - e^{A(2h)}x_o - \int_{-h}^{o} e^{A(h-s)}Cu_o(s)ds \neq 0$$

for all $u(s) \subset R^m$. Let

$$z_u = \int_o^h e^{A(2h-s)}[B + e^{-Ah}C]\, u(s)ds + \int_h^{2h} e^{A(2h-s)}Bu(s)ds$$

and let K be the set of points z_u for all $u(s) \subset R^m$. Then is is easy to see that K is a linear space in R^n. Furthermore, since nonzero vector w is not in K, K is a linear subspace of R^n. Thus there exists a nonzero vector v in R^n which is orthogonal to K. Then

$$vz_u = \int_o^h ve^{A(2h-s)}\,[B + e^{-Ah}C]\, u(s)ds + \int_h^{2h} ve^{A(2h-s)}Bu(s)ds = 0$$

Since this is true for all $u(s) \subset R^m$,

$$ve^{A(2h-s)}B = 0 \qquad \text{for } h \leq s \leq 2h$$

$$ve^{A(2h-s)}e^{-Ah}C = 0 \qquad \text{for } 0 \leq s \leq h$$

and so, for any integer $k \geq 0$

$$\frac{d^k}{ds^k}\, ve^{A(2h-s)}B = (-1)^k vA^k e^{A(2h-s)}B = 0, \text{ for } h \leq s \leq 2h$$

$$\frac{d^k}{ds^k}\, ve^{A(2h-s)}[B+e^{-Ah}C] = (-1)^k vA^k e^{A(2h-s)}[B+e^{-Ah}C] = 0, \text{ for } 0 \leq s \leq h$$

Let $s = h$. Then, for all $k \geq 0$,

$$vA^k e^{Ah}B = ve^{Ah}A^k B = 0$$

$$vA^k e^{Ah}[B+e^{-Ah}C] = vA^k e^{Ah}B + vA^k e^{Ah}e^{-Ah}C = ve^{Ah}A^k e^{-Ah}C = 0$$

Since v is a nonzero vector in R^n, this means that

$$\text{Rank } [e^{Ah}B,\ e^{Ah}AB,\ldots,e^{Ah}A^{n-1}B,\ e^{Ah}e^{-Ah}C,\ e^{Ah}Ae^{-Ah}C,\ldots,e^{Ah}A^{n-1}e^{-Ah}C] < n$$

Now e^{Ah} is nonsingular. Therefore

$$\text{Rank } [B,\ AB,\ldots,A^{n-1}B,\ e^{-Ah}C,\ Ae^{-Ah}C,\ldots,\ A^{n-1}e^{-Ah}C] < n.$$

This is a contradiction. Hence if the rank is n, then the system is controllable.

REFERENCES

1. E. B. Lee and L. Markus, "Foundations of Optimal Control Theory" Wiley, New York, 1967.
2. R. E. Kalman, et al., "Controllability of Linear Dynamical Systems", Contributions to Differential Equations. Vol.1, No.2, 1963.

OBSERVABILITY: ITS ROLE IN CONTROL AND ESTIMATION THEORY

R. E. KLEIN
Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

ABSTRACT

This paper considers the fundamental question of observability. The classes of deterministic systems considered include those modeled by (1) nonlinear ordinary differential equations and (2) partial differential equations. A primary objective is the separation of the observability concept from the properties of the familiar linear state-space model. Attention is directed toward the role of observability in state estimation theory.

INTRODUCTION

Observability and controllability, first introduced by Kalman [1]* in 1960 for linear finite state-space models, are central concepts in the development and analysis of control for deterministic dynamical systems. In the present paper, the notion of observability is formulated so as to apply to both linear and nonlinear and lumped and distributed systems. The concept of observability is fundamental to optimal formulations [2] and, also, the associated problems of filtering and estimation theory [3]. In [1] both observability and controllability were originally defined by Kalman in terms of properties of linear state-space models and the available linear ordinary differential equation theory. The retention of the dependence of observability definitions on properties of linear state-space models has prohibited the orderly extension of this concept to broader classes of systems.

Controllability and open-loop control questions for nonlinear and distributed problems are initial value in nature and solvable, often by numerical techniques, with available mathematical and computational methods. On the other hand, observability results and, as a consequence, closed-loop optimal control are presently unavailable, in general, except for the well-known case of certain finite order linear systems. In fact, only recently has observability been recognized as a central problem in control [4,5].

Observability and state estimation results are lacking for nonlinear systems and many types of distributed parameter dynamical systems. However, certain basic definitions have been formulated in the literature. In [6,7] the author has given a definition of observability for linear and nonlinear distributed parameter systems. Also, Lee and Markus [8] have given a definition of observability *in a local sense* for nonlinear lumped parameter systems.

The linear state-space models originally considered by Kalman have three special properties which are implicitly required for a dynamical model by Kalman's definition of observability. In addition, a fourth system property, that of structural decomposition, was employed by Gilbert [9] and is also of concern. In summary, these properties are

1. linearity of the differential equation operator,
2. group property with respect to time of the transition operator,
3. lumped parameter representation of the model; *i.e.*, absence of a

*Numbers in brackets refer to entries in REFERENCES.

spatial continuum and the associated propagation rates of information in a continuum, and

4. structural decomposition properties of the model (should they exist).

In practice, the first three pose serious restrictions on deterministic models since many accurate models of physical systems may lack any one or possibly all of these three properties. The fourth property concerning structural decomposition is of limited utility. This is evident, for example, in systems described by partial differential equations, where decomposition is usually not possible, *a priori*, due to the multiplicity of independent variables.

The fundamental trait associated with previous observability notions is their inherent dependence on one or more of the above properties. For linear dynamical systems described by finite order ordinary differential equations, usable (and powerful) results have been achieved since theory regarding the four system properties mentioned above is sufficiently complete. However, for certain classes of systems for which observability results are lacking, specifically, for nonlinear systems and distributed parameter dynamical systems, the four properties above are often not present, and, as a consequence, any observability definition which assumes any or possibly all will be over restrictive.

On that basis, if observability results are to be obtained for a wider class of systems, *it is essential that the observability question be rephrased and that the definition of observability itself be divorced from the above system properties and, as an additional consequence, from the familiar state-space model.* Obviously, any new definition should be compatible with earlier versions [1,9] for the case of linear finite systems.

OBSERVABILITY: THE BASIC ISSUES

In considering the application of observability in a closed-loop optimal control scheme where identification of the system dynamics is required, three fundamental issues arise. These are

1. Does the output measurement (assumed to be noise free) contain sufficient information to uniquely prescribe a solution with respect to the dynamical model?

2. Can a stable estimator be built which operates in the presence of noise or data uncertainty?

3. Given an estimator, can something be said about the accuracy of the estimates in the presence of noise?

The uniqueness question is needed as a guarantee that one and only one solution could have produced the observed system output (assuming noise free measurements). Of the distinct solutions which are admissible, conditions must be guaranteed so that a distinct measured output occurs for each of the admissible solutions.

The reconstructibility question arises next (and, logically, only after an affirmative answer to the uniqueness question). Given a particular observed output, is there some efficient and stable method for identifying or estimating the solution which generated the observed output? In essence, then, does there exist any systematic and practical method or device which is capable of processing (or operating on) the observed output to correctly identify or reconstruct the solution over a specified range (or subset) of the independent variables?

The requirement of an error estimate (question 3) is equivalent to requiring that any acceptable estimation procedure must be "well posed" *in the sense of Hadamard* [10]. Broadly speaking, this implies that small errors in an engineering model or its associated data will result only in small errors in the predicted data. This is a natural condition since one must always contend with imperfections in models, measurements, and so forth. This condition, thus, implies the development of suitable *a priori* and/or *a posteriori* error estimates.

Previously, the theory developed for linear finite order systems answered, in particular, both the uniqueness and reconstructability questions simultaneously. Hence, it was not necessary to distinguish these two levels and the two became intermingled. The development of the paper now has led to the central issue: What intrinsic properties constitute observability in a dynamical system? *Observability should be defined exclusively with reference to the uniqueness question* (question 1 above).

With regard to the exclusion of the reconstruction question (question 2 above), the types of computer algorithms, physical devices, circuits, or mechanisms which might eventually be employed to achieve a practical estimator are in no sense unique. If a requirement for observability involved the specification of a reconstruction procedure, this would be equivalent to making a *system property* dependent on state-of-the-art and not on the fundamental and intrinsic properties of the dynamical system.

The exclusion of the accuracy question from the observability question is required so that the definition, to be rendered herein, will remain compatible with the previous and well-known definitions by Kalman [1] and Gilbert [9]. For example, Kalman's definition reduces to the well-known criterion of a rank condition of a matrix when the system is finite, linear, and time invariant. Obviously, a rank condition in no way considers the magnitude of any matrix determinant nor does it in any other way answer the questions related to error analysis and error propagation. It goes, however, as a matter of record, that all three of the above questions, uniqueness, reconstruction, and accuracy, should be answerable (or predictable) in the affirmative before state estimation is complete. Also, the two properties, uniqueness and accuracy, are sufficient to guarantee the *existence* of an estimation procedure.

OBSERVABILITY: A GENERAL DEFINITION

In consideration of the classes of deterministic dynamical systems for which observability and filtering results are sought, the most general system is one which is represented by a nonlinear partial differential equation system such as those derived from the conservation laws of continuum physics [11]. Furthermore, most other types of deterministic dynamical systems of concern are simplified or degenerate forms of the nonlinear distributed parameter case. Therefore, the following definition for observability of nonlinear distributed parameter systems [lumped parameter systems] is given.

Definition 1. Definition of Observability

Given a partial differential equation [ordinary differential equation] model of a deterministic system, the system is said to be observable in a specified space-time domain $\Omega_c(\underline{x},t)$ [time domain $\Omega_c(t)$] if and only if the uniqueness of the dependent variables of the solution of the model is guaranteed in $\Omega_c(\underline{x},t)$ [$\Omega_c(t)$] by the measurement output $y(t)$ generated from another specified space-time measurement domain $\Omega_m(\underline{x},t)$ [time domain $\Omega_m(t)$] where the output transformation generating $y(t)$ is given, along with any other auxiliary conditions or restrictions on the class of admissible solutions.

Remark 1. The definition of observability above is applicable to dynamical systems including those which allow (1) nonlinear operator properties, (2) systems with only semi-group properties with respect to positive time, (3) systems with a continuum spatial domain, (4) no knowledge of structural decomposition properties, and (5) differential delay configurations.

Remark 2. The definition is sufficiently general to encompass the concept of local observability for nonlinear systems in a sense analogous to that of Lee and Markus [8]. This is accomplished by merely restricting appropriately the class of admissible solutions.

Remark 3. It is well known that all systems which are observable in the sense of Kalman [1] are also observable in the sense of Gilbert [9]. Furthermore, all systems which are observable in the sense of Gilbert are observable in the sense of the definition given herein.

Remark 4. The data uncertainties or measurement noises which are associated with a measurement process in no way influence the *observability* of the associated system. Measurement noises will, however, have a bearing on the questions of reconstruction and estimator accuracy.

AN EXAMPLE OF OBSERVABILITY AND STATE ESTIMATION

In another paper [12], Cannon and Klein consider the problem of state estimation in a one-dimensional heat equation given a spatially averaged temperature measurement. The results are summarized herein.

Consider the problem of determining the unknown function $u = u(x,t)$ where u satisfies

$$u_t = u_{xx} , \qquad 0 < x < 1 , \quad 0 < t \leq T \tag{1}$$

$$u_x(0,t) = \phi(t) , \qquad 0 \leq t \leq T$$

$$u_x(1,t) = \psi(t) , \qquad 0 \leq t \leq T \tag{2}$$

$$|u(x,0)| \leq M , \qquad 0 \leq x \leq 1$$

where M and T are given, and such that $u(x,t) \in C^2(\Omega X t) \cap C^1(\bar{\Omega} X t)$ where $\Omega = \{x | 0 < x < 1\}$ and where $\bar{\Omega}$ denotes the boundary of Ω. Even when $\phi(t)$ and $\psi(t)$ are specified exactly, conditions (1),(2) are insufficient to determine $u(x,t)$. Therefore, additional auxiliary conditions are necessary in order to guarantee uniqueness of $u(x,t)$ in $\Omega X t$. Assuming that the process described by (1)-(2) is on-line, consider the use of a measurement $g(t)$ such that

$$g(t) = \int_0^{\delta} k(\xi) u(\xi,t)\, d\xi , \quad 0 < T_o \leq t \leq T_m \leq T , \quad T_o \leq T_m/4$$

$$0 < \delta \leq 1/2 \tag{3}$$

$$0 < k_* \leq k(x) \leq k^* , \quad 0 \leq x \leq \delta$$

where T_o, T_m, k_*, $k(x)$, k^* and δ are given.

Furthermore, assume that $\phi(t)$, $\psi(t)$, and $g(t)$ are noise corrupted and known only approximately as $\phi^*(t)$, $\psi^*(t)$, and $g^*(t)$ such that

$$\sup_{0 \leq t \leq T} |\phi - \phi^*| < \varepsilon_1 \tag{4}$$

$$\sup_{0 \leq t \leq T} |\psi - \psi^*| < \varepsilon_1 \tag{5}$$

$$\sup_{T_o \leq t \leq T_m} |g - g^*| < \varepsilon_2 \tag{6}$$

where ε_1 and ε_2 are known bounds on the data errors.

It is shown in [12] that the observability of the solution is dependent upon $k(x)$ and that $u(x,t)$ is unique (in the sense of each Fourier coefficient a_n) if and only if, for each n,

$$\int_0^{\delta} k(x) \cos n\pi x \, dx \neq 0 \ , \quad n = 0, 1, \cdots \tag{7}$$

Also, *a priori* and *a posteriori* error estimate bounds on $|u(x,\tau) - \hat{u}(x,\tau)|$ are derived as functions of computable constants which may be calculated from the given problem data. Hence, the system given above is reconstructable to a known degree of solution accuracy at some specified time τ even though (7) is not necessarily satisfied for $n \geq N$ where N is a predeterminable number. Lastly, in [12], two independent numerical procedures are given for the generation of an approximate solution $\hat{u}(x,t)$. It is important to note that the reconstruction procedure in [12] requires the storage of the data $\phi^*(t)$, $\psi^*(t)$, and $g^*(t)$. This is not an overly restrictive requirement in a modern engineering system.

The purpose of summarizing the results of [12] is to indicate that the observability concepts and definitions given herein have already led to the successful development of observability results and estimation procedures for systems which, heretofore, were unresolved.

OBSERVABILITY AND ITS RELATION TO ESTIMATION PROCEDURES

Observability serves as a criterion applicable to any system (which is looked at as if the measurements were noise free) to answer in advance the feasibility of state estimation. If a dynamical system is *nonobservable* in the sense defined herein, it is futile to proceed with the building of an estimator. On the other hand, if a system is *observable*, this admits the possibility of a valid estimation procedure. The degree of success will be dependent, in such a system, on an error analysis where the error is measured with respect to some specified norm.

As an example, Thau [13] has derived a set of optimum filtering equations which are supposedly applicable to a class of linear distributed parameter systems which includes, as a specific example, a one-dimensional diffusion equation over the interval $0 \leq x \leq 1$. The diffusion system in [13] may be easily shown to be nonobservable and *ill posed in the sense of Hadamard*. Consequently, the filtering equations in the example derived by Thau will not be asymptotically stable. Asymptotic stability is an indispensable requirement in filtering theory since, for example, any small bias error may lead to large estimation errors. Interestingly enough, the requirement of observability and controllability (in the sense of Kalman) for

stability of optimum filters is well known [14, pp. 363-364] and yet this fact is often ignored or dismissed in contemporary extrapolations of filtering theory. A detailed discussion is forthcoming [15].

Athans [16] states and implies that optimum filters are the only acceptable way to estimate the solutions of distributed parameter systems. This appears to be a precarious position in light of the shortcomings (nonobservability, for example) of [13] which are cited above and, also, due to (1) the developments in the techniques of Cannon and Klein [12] and (2) the extensive body of related literature pertaining to continuation problems in partial differential equations. A partial listing of this literature follows the references.

For the difficult case of optimal nonlinear filtering, a majority of the results heretofore in the literature have consisted of formal manipulations which, due to nonlinearity, lack a sufficient degree of rigor (*i.e.*, observability of these systems is not established). Detchmendy and Sridhar [17] derive sequential estimator equations for a nonlinear system for which it can be demonstrated, by simulation, for example, that a unique solution to the estimator equations does not occur. Hence, the system investigated by Detchmendy and Sridhar is observable, at most, only in a local sense. The interesting account by Fisher [18] fully acknowledges the existence and difficulty of these nonlinear problems.

CONCLUSION

Observability is an intrinsic system property. The concept of observability is useful as a criterion which, if affirmative, admits the possibility of valid state estimation procedures. On the other hand, if a system is nonobservable, the state estimator can in no way yield a meaningful solution.

REFERENCES

1. Kalman, R. E., "On the General Theory of Control Systems," Proc. IFAC (Moscow, 1960), Vol. 1, Butterworths, London (1961).

2. Kalman, R. E., "Contributions to the Theory of Optimal Control," Bol. Soc. Mat. Mex., Vol. 5, pp. 102-119 (1960).

3. Kalman, R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," Jour. Basic Engr. A.S.M.E., Series D, pp. 95-108 (1961).

4. Bucy, R. S., "Recent Results in Linear and Nonlinear Filtering," in Stochastic Problems in Control, published for A.A.C.C. by A.S.M.E., Univ. of Michigan (June 26, 1968).

5. Bucy, R. S., and Joseph, P. D., Filtering for Stochastic Processes with Applications to Guidance, Interscience, New York (1968).

6. Klein, R. E., "An Investigation of Distributed System Observability from a Continuum Physics Viewpoint," Ph.D. Thesis, Purdue University (1969).

7. Goodson, R. E., and Klein, R. E., "A Definition and Some Results for Distributed System Observability," to appear in I.E.E.E. Trans. on Automatic Control (April 1970); see also, Proceedings of the 1969 J.A.C.C., University of Colorado (August 1969).

8. Lee, E. B., and Markus, L., Foundations of Optimal Control Theory, John Wiley and Sons, New York (1967).

9. Gilbert, E. G., "Controllability and Observability in Multivariable Control Systems," J. SIAM on Control, Series A, Vol. 1, No. 2, pp. 128-151 (1963).

10. Hadamard, J., Lectures on Cauchy's Problem in Linear Partial Differential Equations, Dover Publications, New York (1952).

11. Courant, R., and Hilbert, D., Methods of Mathematical Physics, Vol. II, Partial Differential Equations, Interscience, New York (1962).

12. Cannon, J. R., and Klein, R. E., "An Observability Problem for the Heat Equation from Measurement Data," to be submitted.

13. Thau, F. E., "On Optimum Filtering for a Class of Linear Distributed-Parameter Systems," Jour. Basic Engr. A.S.M.E., Series D, pp. 173-178 (1969).

14. Kalman, R. E., "New Methods in Wiener Filtering Theory," in Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability (edited by Bogdanoff, J. L., and Kozin, F.), John Wiley and Sons, New York (1963).

15. Klein, R. E., "Remarks on 'On Optimum Filtering for a Class of Linear Distributed-Parameter Systems'," to be submitted.

16. Athans, M., "Some Remarks on the Control of Distributed Parameter Systems," in Control of Distributed Parameter Systems, published for A.A.C.C. by A.S.M.E., University of Colorado (August 6, 1969).

17. Detchmendy, D. M., and Sridhar, R., "Sequential Estimation of States and Parameters in Noisy Nonlinear Dynamical Systems," Jour. Basic Engr. A.S.M.E., Series D, pp. 362-368 (1966).

18. Fisher, J. R., "Optimal Nonlinear Filtering," in Leondes, C. T. (ed.), Advances in Control Systems, Vol. 5, Academic Press, New York, pp. 197-300 (1967).

RELATED LITERATURE

19. Cannon, J. R., and Douglas, J., Jr., "The Cauchy Problem for the Heat Equation," SIAM J. Num. Anal., Vol. 4, pp. 317-336 (1967).

20. Cannon, J. R., and Douglas, J., Jr., "The Approximation of Harmonic and Parabolic Functions on Half-Spaces from Interior Data," Numerical Analysis of Partial Differential Equations Symposium at Ispra (Varese), 3-11 Luglio 1967, Edizioni Cremonese Roma, pp. 195-230 (1968).

21. Cannon, J. R., "Backward Continuation in Time by Numerical Means of the Solution of the Heat Equation in a Rectangle," M.A. Thesis, Rice University (1960).

22. Cannon, J. R., "Error Estimates for Some Unstable Continuation Problems," J. Soc. Ind. Appl. Math., Vol. 12, No. 2, pp. 270-284 (June 1964).

23. Cannon, J. R., "*A Priori* Estimate for Continuation of the Solution of the Heat Equation in the Space Variable," Annali di Mat. pura ed applicata, Vol. LXVI (66), pp. 155-165 (1964).

24. Cannon, J. R., "A Cauchy Problem for the Heat Equation," Annali di Mat. pura ed applicata, Vol. LXVI (66), pp. 155-165 (1964).

25. Cannon, J. R., and Miller, K., "Some Problems in Numerical Analytic Continuation," J. Soc. Ind. Appl. Math., Numerical Analysis, Vol. 2, No. 1, pp. 87-97.

26. Cannon, J. R., and Knightly, G. H., "The Approximation of the Solution to the Heat Equation in a Half-Strip from Data Specified on the Bounding Characteristics," to appear in SIAM J. on Numerical Analysis.

27. Cannon, J. R., and Cecchi, M. M., "The Numerical Solution of Some Biharmonic Problems by Mathematical Programming Techniques," SIAM J. Num. Anal., Vol. 3, pp. 451-466 (1966).

28. Cannon, J. R., "Some Numerical Results for the Continuation of the Solution of the Heat Equation Backwards in Time," Numerical Solution of Nonlinear Differential Equations, edited by D. Greenspan, John Wiley and Sons, pp. 21-54 (1966).

29. Cannon, J. R., and Knightly, G. H., "Some Continuous Dependence Theorems for Viscous Fluid Motions," to be submitted.

30. Cannon, J. R., and Knightly, G. H., "A Note on Cauchy Problem for the Navier-Stokes Equations," to be submitted.

31. Cannon, J. R., and Klein, R. E., "A Numerical Method for the Solution of a Cauchy Problem for the Heat Equation," to be submitted.

32. Cannon, J. R., and Klein, R. E., "Optimal Selection of Measurement Locations in a Conductor for Approximate Determination of Temperature Distributions," to be submitted.

33. Douglas, J., Jr., "A Numerical Method for Analytic Continuation," Boundary Value Problems in Differential Equations, University of Wisconsin Press, Madison (1960).

34. Douglas, J., Jr., "Mathematical Programming and Integral Equations," Symposium, Provisional International Computation Center, Birkhauser Verlag, Basel (1960).

35. Douglas, J., Jr., and Gallie, T. M., Jr., "An Approximate Solution of an Improper Boundary Value Problem," Duke Math. J., 26, pp. 339-348 (1959).

36. Ford, W. T., "Mathematical Programming and Integro-Differential Equations," J. Soc. Ind. Appl. Math., Series B: Numerical Analysis, Vol. 2, No. 1, pp. 171-202 (1965).

37. Fox, D. W., and Pucci, C., "The Dirichlet Problem for the Wave Equation," Annali di Mat. pura ed applicata (IV), Vol. XLVI, pp. 155-182.

38. Fritz, J., "Numerical Solution of the Equation of Heat Conduction for Preceding Times," Annali di Mat. pura ed applicata 40, pp. 129-142 (1955).

39. Fritz, J., "Continuous Dependence on the Data for Solutions of Partial Differential Equations with a Prescribed Bound," Communications on Pure and Applied Math., Vol. XIII, pp. 551-585 (1960).

40. Ginsberg, F., "On the Cauchy Problem for the One-Dimensional Heat Equation," Math. Comp. 17, pp. 257-269 (1963).

41. Kahane, C., "On the Spatial Analyticity of Weak Solutions of the Navier-Stokes Equations," to appear.

42. Klein, R. E., "State Estimation in a Parabolic Distributed Parameter System," to be submitted.

43. Landis, E. M., "A Three Spheres Theorem," Dokl. Akad. Nauk SSSR 148, pp. 277-279 (1963).

44. Lavrentiev, M. M., "On Cauchy's Problem for Laplace's Equation," Dokl. Akad. Nauk SSSR, Vol. 102, pp. 205-206 (1955).

45. Lavrentiev, M. M., "On the Inverse Problem of Potential Theory," Dokl. Akad. Nauk SSSR, Vol. 106, pp. 389-390 (1956).

46. Miller, K., "Three Circle Theorems in Partial Differential Equations and Applications to Improperly Imposed Problems," Arch. Rat. Mech. Anal. 16, pp. 126-154 (1964).

47. Miller, K., "Stabilized Numerical Analytic Prolongation with Poles," to appear in SIAM J. on Numerical Analysis.

48. Miller, K., "Least Squares Methods for Ill Posed Problems with a Prescribed Bound," to appear.

49. Payne, L. E., "On Some Non-Well Posed Problems for Partial Differential Equations," Numerical Solutions of Nonlinear Differential Equations, edited by D. Greenspan, John Wiley and Sons, pp. 239-262 (1966).

50. Pucci, C., "Sui problemi di Cauchy non 'ben posti'," Rendiconti dell Academia National dei Lincei, Serie 8, Vol. 8, pp. 473-477 (1955).

51. Pucci, C., "Alcune limitazioni per le solutioni de equationi paraboliche," Annali di Mat. pura ed applicata, Serie IV, Tomo XLVIII, pp. 161-172 (1959).

52. Saylor, R. L., "A Generalized Boundary-Integral Representation for Solutions of Elliptic Partial Differential Equations," Ph.D. Thesis, Rice University (1966).

53. Saylor, P. E., "*A Priori* Estimates for Continuation Problems for Elliptic and Principally Normal Differential Equations," Ph.D. Thesis, Rice University (1967).

A MODIFIED LIAPUNOV DESIGN FOR MODEL REFERENCE ADAPTIVE CONTROL SYSTEMS

J. W. GILBART
University of Massachusetts
Amherst, Massachusetts

R. V. MONOPOLI
University of Massachusetts
Amherst, Massachusetts

Abstract -- A modified Liapunov design technique for model reference adaptive control systems is developed. Additional positive semidefinite forms included in a new Liapunov function lead to a modified adaptive law for each adjustable parameter. This modified rule yields comparable system performance with approximately one-sixth of the adaptive loop gain required by the old rule, as well as an improvement in the error response and the transient response for the adaptive parameters.

SUMMARY

Liapunov-designed model reference adaptive control systems introduced in [1] have the advantage over those based on the "M.I.T. rule" in that asymptotic stability of the resulting design is guaranteed. Following its introduction in [1], the design technique was improved and extended by other researchers, [2]-[5]. Presented in the paper is a modification to the general adaptive rule for the class of plants considered in [5]. This class of plants includes the plants considered in [1]-[3], and allows for as many as $m^2 + mr$ adaptive parameters (mth order plant with r inputs). The plant is represented by the state equations (notation as in [5])

$$\dot{\underline{x}}_p = A_p\underline{x}_p + B_p\underline{u} \tag{1}$$

where the state vector $\underline{x}_p = [x_{1p}, x_{2p}, \ldots, x_{mp}]$, the input vector $\underline{u} = [u_1, u_2, \ldots, u_r]$ and A_p and B_p are, respectively, $m \times m$ and $m \times r$ matrices containing adjustable parameters as well as unknown parameters which are constant or slowly time-varying. It is assumed that the desired plant characteristics are represented in a stable model reference system described by

$$\dot{\underline{x}}_m = A_m\underline{x}_m + B_m\underline{u} \tag{2}$$

where the dimensions of $\underline{x}_m$, A_m, and B_m are the same as their counterparts in (1) and A_m and B_m are known constant matrices. The design procedure may be summarized as follows. A potential Liapunov function is chosen, which is given by

$$V = \underline{e}^T Q\underline{e} + \sum_{i,j=1}^{m} \frac{1}{\alpha_{ij}}\left[a_{ij} + \beta_{ij}\sum_{k=1}^{m} e_k q_{ki} a_{ij} x_{jp}\right]^2 + \sum_{i,j=1}^{m} \frac{1}{\gamma_{ij}}\left[b_{ij} + \delta_{ij}\sum_{k=1}^{m} e_k q_{ki} b_{ij} u_j\right]^2 \tag{3}$$

where $\underline{e} = \underline{x}_m - \underline{x}_p$ and Q is a symmetric matrix to be determined with elements q_{ij}; a_{ij} and b_{ij} are the elements of $A = A_m - A_p$ and $B = B_m - B_p$, respectively, and $\alpha_{ij}, \beta_{ij}, \gamma_{ij}, \delta_{ij} > 0$. The two summation expressions in each squared term of (3) are new here and are not included in the Liapunov function of [5]. The adaptive law that results from the choice of (3) consists of a particular time integral, which in previous designs constituted the entire adaption law, plus the integrand of this time integral. It is shown that this modified adaptive rule yields comparable system performance

with about one-sixth of the previously required adaptive loop gain, while requiring only one additional summing amplifier. It is also shown that the error response and the transient response of the adaptive parameters are improved. The results of an analog simulation of a second-order system with adaptive gain; i.e., $K_c(t)$, are presented in Fig. 1 to illustrate these advantages. The adaptive loop gains in the previous and modified designs; i.e., 93 and 16, were chosen so that the integral absolute error

$\left(\int_0^\infty |e(t)|\, dt = 0.63 \right)$ was the same for both designs.

REFERENCES

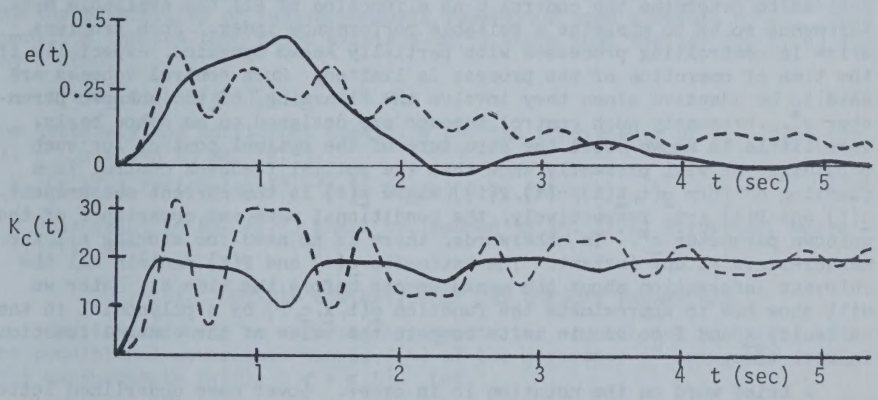
[1] Parks, P. C., "Liapunov Redesign of Model Reference Adaptive Control Systems," IEEE Transactions on Automatic Control, July 1966, Vol. AC-11, No. 3, pp. 362-367.

[2] Monopoli, R. V., "Liapunov's Method for Adaptive Control System Design," IEEE Transactions on Automatic Control, June 1967, Vol. AC-12, No. 3, pp. 334-335.

[3] Phillipson, P. H., "Concerning Liapunov Redesign of Model Reference Adaptive Control Systems," IEEE Transactions on Automatic Control, October 1967, Vol. AC-12, No. 5, p. 625.

[4] Monopoli, R. V., Gilbart, J. W., and Thayer, W. D., "Model Reference Adaptive Control Based on Liapunov-Like Techniques," Proceedings of the Second IFAC Symposium on System Sensitivity and Adaptivity, pp. F.24 - F.31, August 1968.

[5] Winsor, C. A. and Roy, R. J., "Design of Model Reference Adaptive Control Systems by Liapunov's Second Method," IEEE Transactions on Automatic Control, April 1968, Vol. AC-13, No. 2, p. 204.



– – – – Previous Design: $K_c(t) = 93 \int_0^t (eu + \dot{e}u)\, dt$

——— Modified Design: $K_c(t) = 16 \left[eu + \dot{e}u + \int_0^t (eu + \dot{e}u)\, dt \right]$

Fig. 1: Time Response of Second Order System

OPTIMAL CONTROL OF A CLASS OF STOCHASTIC SYSTEMS WITH UNKNOWN PARAMETERS*

R. L. KASHYAP
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

We consider the determination of the optimal feedback control function for a system described by a stochastic differential equation with some of the parameters unknown. The state of the system is assumed to be exactly measurable. We will first derive recursive equations for the conditional mean and covariance of the unknown parameters. We will show that the optimal feedback control is a function of the current measurement, the conditional mean and variance of the unknown parameters. We mention some methods for the approximation of the optimal feedback control function, and the results of simulation of the control scheme.

I. INTRODUCTION

Consider the stochastic control system described by the vector stochastic differential equation (1.1)

$$\dot{\underline{x}} = \underline{a}\,(\underline{x}(t),t) + B(\underline{x}(t),\underline{t}))\ \underline{c}^* + G(t)u(t) + F(t)\ \dot{\underline{\omega}} \qquad (1.1)$$

where the vector function $\underline{a}(x,t)$ and the matrix functions $B(\underline{x},t)$, $G(t)$ and $F(t)$ are known, $\underline{\omega}$ is the white gaussian noise and $\underline{c}^*$ is a constant vector whose true value is unknown. $\underline{c}^*$ may represent the unknown parameters of the system $\underline{a}(\underline{x},t)$ or it may represent the unknown mean of the disturbance $\dot{\underline{\omega}}(t)$. The state vector $\underline{x}(t)$ is amenable for continuous measurement. It is desired to determine the control $\underline{u}$ as a function of all the available measurements so as to minimize a suitable performance index. Such problems arise in controlling processes with partially known dynamics, especially if the time of operation of the process is limited. Such control schemes are said to be adaptive since they involve the "learning" of the unknown parameter $\underline{c}^*$. Presently such control schemes are designed on an adhoc basis. Very little is known about the structure of the optimal control for such problems. We will presently show that the optimal feedback control is a function of form $\underline{\varphi}(t,\underline{x}(t),c(t),P(t))$ where $\underline{x}(t)$ is the current measurement, $\underline{c}(t)$ and $P(t)$ are, respectively, the conditional mean and covariance of the unknown parameter $\underline{c}^*$. In otherwords, there is no need for storing all the measurements at any instant. The estimates $\underline{c}(t)$ and $P(t)$ contain all the relevant information about the measurements before the time t. Later we will show how to approximate the function $\underline{\varphi}(t,\underline{x},\underline{c},P)$ by a polynomial in the va iables $\underline{x}$ and P to enable us to compute the value of the control function in real time.

A brief word on the notation is in order. Lower case underlined letters stand for vectors, upper case underlined letters represent matrices. The transpose is indicated by a prime. Scalars will be mentioned explicitly. The (i,j)th element of the matrix A is indicated by $(A)_{ij}$. A matrix $R > 0$ means that R is positive definite. If $f(\underline{x},\underline{c})$ is a scalar function of vectors $\underline{x}$ and $\underline{c}$, $(\partial^2 f/\partial\underline{x}\partial\underline{c})$ represents a matrix whose (i,j)th element is $(\partial^2 f/\partial\underline{x}_i\partial\underline{c}_j)$.

II. STATEMENT OF THE PROBLEM

It is convenient to rewrite the Eq. (1.1) in incremental form

$$d\underline{x}(t) = a(\underline{x}(t),t)dt + B(\underline{x}(t),t)\underline{c}^*dt + G(t)\underline{u}(t)dt + F(t)d\underline{\omega}(t) \qquad (?.1)$$

where $\underline{x}(t)$ is the state vector of dimension n, $\underline{c}^*$ is an unknown parameter

vector of dimension n_1, $\underline{a}(\underline{x},t)$ is a known vector function of dimension n, $B(\underline{x},t)$ is a known matrix function of dimension nxn_1, $\underline{u}(t)$ is a control vector of dimension n_2, $G(t)$ is a known matrix of dimension nxn_2, $F(t)$ is a known matrix of dimension nxn_3, $n_3 \leq n$, and $\underline{w}(t)$ is an n_3-dimensional Wiener Process with the following properties:

$$E(\underline{w}(t)) = 0\ , \quad E(\underline{w}(t)\underline{w}'(s)) = Q_1 \min(t,s), \quad Q_1 > 0, \quad (Q_1)_{ij}=0 \text{ if } i\neq j. \tag{2.2}$$

It is required to find a feedback control

$$\underline{u}(t) = \underline{\varphi}(t,\underline{x}(v);\ 0 \leq v \leq t) \tag{2.3}$$

so as to minimize the criterion function $J(\underline{u})$

$$J(\underline{u}) = \frac{1}{2}\, E \int_0^T \left(\|\underline{h}(\underline{x}(t))\|^2 + \|\underline{u}(t)\|^2_{R_2} \right) dt\ . \tag{2.4}$$

It is useful to regard the unknown constant parameter $\underline{c}^*$ as a random vector whose realization remains a constant throughout the experiment. Its a priori density is given in Eq. (2.5).

$$\underline{c}^* \sim N(\underline{c}(0), P(0)) \tag{2.5}$$

To present the solution, we need the equation for the optimal estimate $\underline{c}(t)$ of $\underline{c}^*$ which is given in the next section. Before doing this, we observe from Eq. (2.1) that it may be possible to estimate accurately some of the components of $\underline{c}^*$ instantaneously. Since this situation is unrealistic, we have to impose additional conditions on the system matrices B and F to avoid its occurence. The condition can be found by inspection of Eq. (2.1). In effect, it says that in all those component equations of (2.1) in which at least one element of $\underline{c}^*$ occurs explicitly, at least one element of $d\underline{w}(t)$ must also occur explicitly. The appearance of the noise components prohibits the possibility of instantaneously recovering some elements of $\underline{c}^*$. This condition implies that there exists a $d\underline{w}(t)$ which must satisfy Eq. (2.6) for any $\underline{c}^*dt=0$.

$$F(t)d\underline{w}(t) + B(\underline{x}(t),t)\ \underline{c}^*dt = 0\ . \tag{2.6}$$

The necessary and sufficient condition for the consistency of Eq. (2.6) for all $\underline{c}^*$ is given in Eq. (2.7).

$$\left(I-F(t)\ (F'(t)F(t))^{-1}\ F'(t)\right) B(\underline{x},t) = 0, \text{ for all } \underline{x}\ . \tag{2.7}$$

The constraint in Eq. (2.7) shall be imposed in Eq. (2.1) hereafter, where $F'(t)F(t) > 0$.

III. OPTIMAL ESTIMATION OF THE UNKNOWN PARAMETERS

Our intention is to recursively compute the conditional mean $\underline{c}(t)$ and the conditional covariance matrix $P(t)$ of the paremeter $\underline{c}^*$ using the available measurements $\{\underline{x}(v);\ 0 \leq v \leq t\}$. Let

$$\underline{c}(t) = E[\underline{c}^* | x(v),\ 0 \leq v \leq t]$$

$$P(t) = E[(\underline{c}^*-\underline{c}(t))\ (\underline{c}^*-\underline{c}(t))' | \underline{x}(v),\ 0 \leq v \leq t]\ .$$

Let us rewrite Eq. (2.1) after left-multiplying it by the matrix F'

$$d\underline{z}(t) = \overline{B}(\underline{x}(t),t)\underline{c}^*dt + F'(t)F(t)d\omega(t) \tag{3.1}$$

where

$$d\underline{z}(t) \triangleq F'(t)d\underline{x}(t) - F'(t)\underline{a}(\underline{x}(t),t)dt - F'(t)G(t)\underline{u}(t)dt$$

$$\overline{B}(\underline{x},t) = F'(t)B(x,t) \tag{3.2}$$

Let

$$Q(t) \triangleq F'(t)F(t)Q_1F'(t)F\ (t) > 0 \tag{3.3}$$

Using Eq. (3.1) it is easy to write the conditional density of $\underline{c}^*$ given the measurements $\{\underline{x}(v), 0 \le v \le t\}$. To do this, we will divide the time interval $[0,t]$ into N increments of width Δ so that $N\Delta = t$. Let the corresponding approximations of the processes $\underline{x}(t)$, $\underline{w}(t)$ and $\underline{z}(t)$ be $\bar{\underline{x}}(N\Delta)$, $\bar{\underline{w}}(N\Delta)$ and $\bar{\underline{z}}(N\Delta)$. Let

$$d\bar{\underline{w}}(i\Delta) = \bar{\underline{w}}\,(\overline{i+1}\Delta) - \bar{\underline{w}}\,(i\Delta)$$

$$d\bar{\underline{x}}(i\Delta) = \bar{\underline{x}}\,(\overline{i+1}\Delta) - \bar{\underline{x}}\,(i\Delta)$$

$$d\bar{\underline{z}}(i\Delta) = F'(i\Delta)\,\{d\bar{\underline{x}}(i\Delta) - \underline{a}(\underline{x}(i\Delta), i\Delta)\Delta - G(i\Delta)\underline{u}(i\Delta)\Delta\}$$

$$p(\underline{c}^* | d\bar{\underline{z}}(0), d\bar{\underline{z}}(\Delta), ., d\bar{\underline{z}}(\overline{N-1}\Delta)) = k \exp\Big[-\frac{1}{2\Delta}\sum_{i=0}^{N-1} \|d\bar{\underline{z}}(i\Delta) -$$

$$\bar{B}(\bar{\underline{x}}(i\Delta), i\Delta)\underline{c}^*\Delta\|^2_{Q^{-1}(i\Delta)} - \frac{1}{2}\|\underline{c}^* - \underline{c}(0)\|^2_{P^{-1}(0)}\Big]$$

$$= k_1 \exp\Big[-\frac{1}{2}\|\underline{c}^* - \bar{\underline{c}}(N\Delta)\|^2_{(\bar{P}(N\Delta))^{-1}}\Big] \tag{3.4}$$

where k_1 is a normalizing constant

$$(\bar{P}(N\Delta))^{-1} = \sum_{i=0}^{N-1} \Delta\bar{B}'(\bar{\underline{x}}(i\Delta), i\Delta)Q^{-1}(i\Delta)\bar{B}(\bar{\underline{x}}(i\Delta), i\Delta) + P^{-1}(0) \tag{3.5}$$

$$\bar{\underline{c}}(N\Delta) = \bar{P}(N\Delta)\Big\{\sum_{i=0}^{N-1} \bar{B}'(\bar{\underline{x}}(i\Delta), i\Delta)Q^{-1}(i\Delta)d\bar{\underline{z}}(i\Delta) + P^{-1}(0)c(0)\Big\} \tag{3.6}$$

In Eq. (3.5), take the limit as $N \to \infty$, $\Delta \to 0$ so that $N\Delta = t$ and $\bar{P}(N\Delta)$, $\bar{\underline{x}}(N\Delta)$ tend to $P(t)$ and $\underline{x}(t)$, respectively.

$$P^{-1}(t) = \int_0^t \bar{B}'(\underline{x}(u), u)Q^{-1}(u)\bar{B}(\underline{x}(u), u)du + P^{-1}(0). \tag{3.7}$$

We can express (3.7) as a differential equation in $P(t)$.

$$\Big(dP(t)/dt\Big) = -P(t)B'(x(t), t)F(t)Q^{-1}(t)F'(t)B(\underline{x}(t), t)P(t) \tag{3.8}$$

To consider the limit of $\bar{\underline{c}}(N\Delta)$, let us rewrite Eq. (3.5) using Eq. (3.1)

$$\bar{\underline{c}}(N\Delta) = \bar{P}(N\Delta)\,\{P^{-1}(0)\underline{c}(0) + \sum_{i=0}^{N-1} (\bar{\underline{x}}(i\Delta), i\Delta)Q^{-1}(i\Delta)\bar{B}(\bar{\underline{x}}(i\Delta), i\Delta)\underline{c}^*\Delta$$

$$+ \sum_{i=0}^{N-1} \bar{B}'(\bar{\underline{x}}(i\Delta), i\Delta)Q^{-1}(i\Delta)F'(i\Delta)F(i\Delta)d\bar{w}(i\Delta)\} \tag{3.9}$$

Since we do not know the limit of the last term in Eq. (3.9) as $\bar{w}(i\Delta)$ tends to a Wiener process by inspection, the limiting operation is done in two steps. In the first step, $\bar{w}(N\Delta)$ tends to $\bar{w}(t)$ where $\bar{w}(t)$ is a wideband process. Correspondingly,

$$\bar{\underline{c}}(N\Delta) \to \bar{\underline{c}}(t), \quad \bar{\underline{x}}(N\Delta) \to \bar{\underline{x}}(t)$$

where $\bar{\underline{x}}(t)$ obeys an equation like (2.1) with $dw(t)$ replaced by $d\bar{w}(t)$. Since $d\bar{w}(t)$ is well defined, Eq. (3.9) yields

$$\bar{\underline{c}}(t) = P(t)\Big\{P^{-1}(0)\underline{c}(0) + \int_0^t \bar{B}'(\bar{\underline{x}}(v), v)Q^{-1}(v)\bar{B}(\bar{\underline{x}}(v), v)\underline{c}^* dt$$

$$+ \int_0^t \bar{B}'(\bar{\underline{x}}(v),v)Q^{-1}(v)F'(v)F(v)d\bar{\underline{w}}(v)\Big\} \qquad (3.11)$$

Differentiating (3.11) with respect to t and using (3.8), we get

$$d\bar{c}(t) = (dP(t))P^{-1}(t)\bar{c}(t) + P(t)\left\{\bar{B}'(\bar{\underline{x}}(t),t)Q^{-1}(t)\bar{B}(\bar{\underline{x}}(t),t)\underline{c}^*dt + \bar{B}(\bar{\underline{x}}(t),t)Q^{-1}(t)F'(t)F(t)d\bar{\underline{w}}(t)\right\}$$

$$= P(t)\bar{B}'(\bar{\underline{x}}(t),t)Q^{-1}(t)\bar{B}(\bar{\underline{x}}(t),t)\ (\underline{c}^*-\bar{\underline{c}}(t))dt + P(t)\bar{B}'(\bar{\underline{x}}(t),t)Q^{-1}(t)F'(t)F(t)d\bar{\underline{w}}(t) \qquad (3.12)$$

Now, we will let $\bar{\underline{w}}(t)$ tend to the Wiener process $\underline{w}(t)$ defined in Eqs. (2.2) and (2.3). Correspondingly, $\bar{\underline{x}}(t)$ tends to $\underline{x}(t)$ obeying Eq. (2.1) and $\bar{\underline{c}}(t)$ tends to the conditional mean $\underline{\hat{c}}(t)$ using the results of Wong and Zakai[4,2].

$$d\underline{\hat{c}}(t) = P(t)\bar{B}'(x(t),t)Q^{-1}(t)\left\{\bar{B}(\underline{x}(t),t)\ (\underline{c}^*-\underline{\hat{c}}(t))dt + F'(t)F(t)d\underline{w}(t)\right\} + \underline{f}(x(\underline{t}),P(t),t)dt \qquad (3.13)$$

where

$$f_i(\underline{x},P,t) = \frac{1}{2}\sum_{j,k,\ell}\frac{\partial}{\partial x_k}\left\{(PB'(\underline{x},t)F(t)Q^{-1}(t)F'(t)F(t))_{ij}\right\}F_{k\ell}(t)(Q_1)_{\ell j}$$

$$= \sum_{\ell} P_{i\ell}\,b_\ell(\underline{x},t)$$

where

$$b_\ell(\underline{x},t) \triangleq \frac{1}{2}\sum_{k=1}^{n}(\partial B_{\ell k}(\underline{x},t)/\partial x_k), \quad \ell = 1,\ldots,n_1 \qquad (3.14)$$

We can further simplify the Eq. (3.13) by introducing a new process $\underline{y}(t)$ whose increment $d\underline{y}(t)$ is defined below

$$d\underline{y}(t) \triangleq F'(t)\left\{d\underline{x}(t) - \underline{a}(\underline{x}(t),t)dt - B(\underline{x}(t),t)\underline{c}(t)dt - G(t)\underline{u}(t)dt\right\} \qquad (3.16)$$

Eqs. (3.16) and (2.1) give us the following expression for $d\underline{y}(t)$:

$$d\underline{y}(t) = F'(t)B(\underline{x}(t),t)\ (\underline{c}^*-\underline{c}(t))dt + F'(t)F(t)d\underline{w}(t) \qquad (3.17)$$

Substitution of (3.17) and (3.15) in Eq. (3.13) gives us the following stochastic differential equation for $\underline{c}(t)$

$$d\underline{c}(t) = P(t)B'(\underline{x}(t),t)F(t)Q^{-1}(t)d\underline{y}(t) + P(t)\underline{b}(\underline{x}(t),t)dt\ . \qquad (3.18)$$

Note from Eq. (3.16) that the $\underline{y}(.)$ process is amenable for direct measurement. Thus, Eqs. (3.18) and (3.8) are the required differential equations used for computing the conditional mean and variance of $\underline{c}^*$.

The term $P(t)\underline{b}(\underline{x},t)dt$ in Eq. (3.18) should be mentioned. Since $\underline{c}^*$ is a constant, one would expect $d\underline{c}(t)$ to involve only the first term in R.H.S. of Eq. (3.18) involving $d\underline{y}(t)$. This would be the case if we obtained the equation for $\underline{c}(t)$ using Kalman-Bucy [5] theory formally. But, in Eq. (3.18) there is another term, $P(t)b(x,t)dt$, which comes in on account of the special properties of the white noise process. This term should go to zero as t goes to infinite since P(t) goes to zero with probability one as t goes to infinity.

Let us consider the incremental stochastic process $d\underline{y}(t)$, the so-called innovation process. We can show that the process $d\underline{y}(t)$ has the same first and second order properties as the process $F'(t)F(t)d\underline{w}(t)$ as

mentioned in Eq. (4.4). This is a special case of the general result obtained by Frost and Kailath mentioned in [2]. This fact is of crucial importance in our theory.

IV. OPTIMIZATION SCHEME

All the information available about the unknown $\underline{c}^*$ at time t is contained in the quantities $\underline{c}(t)$ and P(t) which are observable. Thus, we can rewrite the differential equation for $\underline{x}(t)$ by eliminating $\underline{c}^*$ from Eqs. (3.17), (2.7) and (2.1).

$$d\underline{x}(t) = a(\underline{x}(t),t)dt + B(\underline{x}(t),t)\underline{c}(t)dt + G(t)\underline{u}(t)dt + F(t)\,(F'(t)F(t))^{-1}\,d\underline{y}(t) \tag{4.1}$$

The equations for $d\underline{c}(t)$ and dP(t) are given below:

$$d\underline{c}(t) = P(t)B'(\underline{x}(t),t)F(t)Q^{-1}(t)d\underline{y}(t) + P(t)b(\underline{x}(t),t)dt \tag{4.2}$$

$$dP(t) = -\,P(t)B'(\underline{x}(t),t)F(t)Q^{-1}(t)F'(t)B(\underline{x}(t),t)P(t)dt \tag{4.3}$$

The incremental process $d\underline{y}(t)$ (which is exactly observable) is the random forcing term in Eqs. (4.3)-(4.5). The properties of the process $d\underline{y}(t)$ are indicated in Eq. (4.4).

$$\begin{aligned} &E(d\underline{y}(t)) = 0\,, \quad E(d\underline{y}(t)\underline{x}'(t)) = 0 \\ &E(d\underline{y}(t)d\underline{y}'(t)) = Q(t)dt\,, \quad E(d\underline{y}(t_1)d\underline{y}'(t_2)) = 0(dt) \end{aligned} \tag{4.4}$$

where the intervals (t_1, t_1+dt) and (t_2, t_2+dt) are non-overlapping.

We have to find a feedback control u(.) to minimize J(u) subject to the differential constraints in Eqs. (4.1)-(4.3) with $R_2 > 0$.

$$J(u) = \frac{1}{2}\,E\int_0^T \left(\|\underline{h}(\underline{x}(t))\|^2 + \|\underline{u}(t)\|^2_{R_2}\right) dt \tag{4.5}$$

Since all the entities in Eqs. (4.1)-(4.3) are either (stochastic) state variables or known parameters, we can use the principle of optimality to obtain the solution. The procedure is similar to the one adopted by Florentin [3]. We need to define the optimal cost function $V(t,\underline{x},\underline{c},P)$. Let

$$V(t,\underline{x}^*,\underline{c}^*,P^*) = \min_{\substack{u(v) \\ t \le v \le T}} \left[\frac{1}{2}\,E\int_t^T \left(\|\underline{h}(\underline{x}(v))\|^2 + \|\underline{u}(v)\|^2_{R_2}\right) dv\right]$$

with

$$\underline{x}(t) = \underline{x}^*, \underline{c}(t) = \underline{c}^*,\ P(t) = P^*.$$

By the principle of optimality, we have

$$V(t,\underline{x},\underline{c},P) = \min_{u(t)}\ \lim_{dt\downarrow 0}\ E\left\{\left(V\left(t+dt, \underline{x}+d\underline{x}(t), \underline{c}+d\underline{c}(t), P+dP(t)\right) + \left(\frac{1}{2}\,\|\underline{h}(\underline{x}(t))\|^2 + \frac{1}{2}\,\|u(t)\|^2_{R_2}\right)dt\right) \Big| \underline{x}(t) = \underline{x}, \underline{c}(t) = \underline{c}, P(t) = P\right\} \tag{4.6}$$

Expanding the optimal cost function V on the right hand side of Eq. (4.6) about the point $(t,\underline{x},\underline{c},P)$ up to second order terms, substituting for the quantities $d\underline{x}(t)$, $d\underline{c}(t)$ and dP(t) from Eq. (4.1)-(4.3), and taking expectations by using the properties in Eq. (4.4), we obtain the following partial differential equation for $V(t,\underline{x},\underline{c},P)$:

$$\frac{-\partial V}{\partial t} = \min_{u(t)} \left[\left(\frac{\partial V}{\partial \underline{x}}\right)'\left(\underline{a}(\underline{x},t)+B(\underline{x},t)\underline{c}+G(t)\underline{u}(t)\right) + \left(\frac{\partial V}{\partial \underline{c}}\right)' P\underline{b}(\underline{x},t) + \text{Trace}\left\{-\frac{\partial V}{\partial P}\cdot PB'(\underline{x},t)F(t)Q^{-1}(t)F'(t)B(\underline{x},t)P\right.\right.$$

$$+ \frac{1}{2}\frac{\partial^2 V}{\partial \underline{x}^2}\cdot F(t)Q_1F'(t) + \frac{1}{2}\frac{\partial^2 V}{\partial \underline{c}^2}\cdot PB'(\underline{x},t)F(t)Q^{-1}(t)F'(t)B(\underline{x},t)P$$

$$+ \frac{\partial^2 V}{\partial \underline{x}\partial \underline{c}}\cdot PB'(\underline{x},t)F(t)\Big(F'(t)F(t)\Big)^{-1}F'(t)\Big\} + \frac{1}{2}\|\underline{h}(\underline{x})\|^2 + \frac{1}{2}\|u(t)\|^2_{R_2}\Big] \quad (4.7)$$

Minimizing the right hand side of Eq. (4.7) over $\underline{u}(t)$, one gets the following expression for the optimal control

$$\underline{u}^o(t) = -R_2^{-1}G'(t)\,\frac{\partial V(t,\underline{x},\underline{c},P)}{\partial \underline{x}}\Bigg|_{\underline{x}=\underline{x}(t),\,\underline{c}=\underline{c}(t),\,P=P(t)} \quad (4.8)$$

Substitution of Eq. (4.8) in Eq. (4.7) gives the required PDE for the cost function $V(t,\underline{x},\underline{c},P)$.

$$\frac{-\partial V}{\partial t} = \frac{1}{2}\|\underline{h}(\underline{x})\|^2 - \frac{1}{2}\left\|\frac{\partial V}{\partial \underline{x}}\right\|^2_{G(t)R_2^{-1}G'(t)} + \left(\frac{\partial V}{\partial \underline{x}}\right)'\Big(\underline{a}(\underline{x},t)+B(\underline{x},t)\underline{c}\Big)+\left(\frac{\partial V}{\partial \underline{c}}\right)'P\underline{b}(\underline{x},t)$$

$$+ \text{Trace}\Big\{-\frac{\partial V}{\partial P}\cdot PB'(\underline{x},t)F(t)Q^{-1}(t)F'(t)B(\underline{x},t)P + \frac{1}{2}\frac{\partial^2 V}{\partial \underline{x}^2}F(t)Q_1F'(t)$$

$$+ \frac{\partial^2 V}{\partial \underline{x}\partial \underline{c}}\cdot PB'(\underline{x},t)F(t)\Big(F'(t)F(t)\Big)^{-1}F'(t) + \frac{1}{2}\frac{\partial^2 V}{\partial \underline{c}^2}\cdot PB'(\underline{x},t)F(t)Q^{-1}(t)\cdot$$

$$F'(t)B(\underline{x},t)P\Big\} \quad (4.9)$$

The boundary condition for the partial differential Eq. (4.9) is $V(T,\underline{x},\underline{c},P) = 0$.

Eq. (4.8) indicates that the $\underline{u}^o(t)$, the optimal feedback control at instant t is a function of $\underline{x}(t)$, the conditional mean $\underline{c}(t)$ and conditional covariance $P(t)$ alone. This implies that there is no need for storing the entire measurement history $\{\underline{x}(v),\ 0 \le v \le t\}$.

The actual evaluation of the optimal feedback control function depends on the evaluation of the optimal cost function $V(t,\underline{x},\underline{c},P)$ obeying the PDE (4.9), which is not easy to solve.

V. EXAMPLE: FIRST ORDER SYSTEM WITH UNKNOWN TIME CONSTANT

The system equation is given below:

$$dx(t) = c^*x(t)dt + u(t)dt + d\omega(t) \quad (5.1)$$

where the parameter c^* is constant, but unknown and $d\omega(t)$ is the increment of the Wiener Process with zero mean and covariance dt. Let the a priori distribution of $\bar{c}$ be $N(c(0),p(0))$. The criterion function $\bar{J}(u)$ is

$$\bar{J}(u) = E\,[J(u)];\ J(u) = \frac{1}{2}x^2(T) + \frac{1}{2}\int_0^T u^2(t)dt\ .$$

We wish to find the optimal feedback control $u^o(\cdot)$. To do this, we need the equations for $c(t),p(t)$. The conditional mean and variance of $\bar{c}$ are given below:

$$dc(t) = p(t)x(t)dy(t) + \frac{1}{2}p(t)dt, \quad c(0) \text{ is given}$$

$$dp(t) = -p^2(t)x^2(t)dt, \quad p(0) \text{ is given} \quad (5.2)$$

where

$$dy(t) = dx(t) - c(t)x(t)dt - u(t)dt.$$

The optimal control $u^o(t)$ is given by

$$u^o(t) = - \left.\frac{\partial V(t,x,c,p)}{\partial x}\right|_{x=x(t),\ c=c(t),\ p=p(t)} \tag{5.3}$$

The PDE for the optimal cost function $V(t,x,c,p)$ is given below

$$- \frac{\partial V}{\partial t} = - \frac{1}{2}\left(\frac{\partial V}{\partial x}\right)^2 + \left(\frac{\partial V}{\partial x}\right)cx + \frac{1}{2}\left(\frac{\partial V}{\partial c}\right)p - \left(\frac{\partial V}{\partial p}\right)p^2x^2 + \frac{1}{2}\frac{\partial^2 V}{\partial x^2} + \frac{1}{2}\frac{\partial^2 V}{\partial c^2}p^2x^2 + \frac{\partial^2 V}{\partial x \partial c}px \tag{5.4}$$

The boundary condition is: $V(T,x,c,p) = \frac{1}{2}x^2$.

In order to solve the PDE (5.4), we notice that the covariance $p(t)$ decreases with time; and, hence, we can approximate the solution of Eq. (5.4) to any accuracy by the following power series in p

$$V(t,x,c,p) = f_1(t,x,c) + f_2(t,x,c)p + f_3(t,x,c)p^2 + \ldots \tag{5.5}$$

We retain only the first two terms of this series. Substituting Eq. (5.5) in Eq. (5.4) and equating the coefficients of the terms p^o, p^1 on either side, we get the following PDE's in only 3 variables, x, c, and t.

$$- \frac{\partial f_1}{\partial t} = - \frac{1}{2}\left(\frac{\partial f_1}{\partial x}\right)^2 + \left(\frac{\partial f_1}{\partial x}\right)cx + \frac{1}{2}\frac{\partial^2 f_1}{\partial x^2} \tag{5.6}$$

$$- \frac{\partial f_2}{\partial t} = - \frac{\partial f_1}{\partial x}\frac{\partial f_2}{\partial x} + \frac{\partial f_2}{\partial x}cx + \frac{1}{2}\frac{\partial^2 f_2}{\partial x^2} + \frac{\partial^2 f_1}{\partial x \partial c}x + \frac{1}{2}\frac{\partial f_1}{\partial c} \tag{5.7}$$

The boundary conditions for Eqs. (5.6) and (5.7) are

$$f_1(T,x,c,p) = \frac{1}{2}x^2, \quad f_2(T,x,c,p) = 0 \tag{5.6}'$$

To solve the PDE (5.6) with boundary condition in Eq. (5.6)', let us try the following candidate for solution:

$$f_1(t,x,c) = \frac{1}{2}g_1(t,c)x^2 + f_4(t,c) \tag{5.8}$$

Substituting (5.8) in (5.6) and equating the coefficient of x^2 to zero, we get the following differential equation for $g_1(t,c)$ with $g_1(T,c)=1.0$:

$$\frac{\partial g_1(t,c)}{\partial t} + 2cg_1(t,c) - g_1^2(t,c) = 0. \tag{5.9}$$

Eq. (5.9) has the following closed form solution:

$$g_1(t,c) = \exp[2c(T-t)] \,/\, [1 - (1-\exp(2c(T-t))) \,/\, 2c].$$

Similarly, the PDE (5.7) has the following solution:

$$f_2(t,x,c) = \frac{1}{2}x^2 g_2(t,c) + f_5(t,c) \tag{5.10}$$

where $g_2(t,c)$ obeys the following PDE with $g_2(T,c) = 0$:

$$- \frac{1}{2}\frac{\partial g_2}{\partial t} = - g_1 g_2 + g_2 c + \frac{5}{4}\frac{\partial g_1}{\partial c}. \tag{5.11}$$

For any given value of c, we can find the values of $g_2(t,c)$ for all t by treating Eq. (5.11) as an ordinary differential equation in t.

Retaining only the first 2 terms in the expansion (5.5) and using Eq. (5.3), we can approximate the optimal control by the following expression :

$$u^*(t) = - \left(\frac{\exp[2c(t)(T-t)]}{\left\{1 - \frac{1-\exp[2c(t)(T-t)]}{2c(t)}\right\}} + p(t)g_2(t,c(t)) \right) \underline{x}(t). \tag{5.14}$$

where $c(t)$ and $p(t)$ are obtained by integrating Eq. (5.2) in forward

direction and $g_2(t, c(t))$ is obtained by integrating Eq. (5.11) in backward direction.

In order to get an idea of the effectiveness of the control scheme in Eq. (5.14), we will compare it with 2 other schemes, $\bar{u}(t)$ and $u^{**}(t)$. $\bar{u}(t)$ is the optimal control if $\bar{c}$ were known.

$$\bar{u}(t) = - \frac{\exp 2c^*(T-t)]}{\left\{1 - \frac{1-\exp(2c^*(T-t))}{2c^*}\right\}} \; x(t) \tag{5.15}$$

$u^{**}(t)$ is the control scheme obtained by replacing c^* by $c(0)$ in Eq. (5.15). This is the control scheme one would use if we did not have the method of this paper.

$$u^{**}(t) = - \frac{\exp(2c(0)(T-t)}{\left\{1 - \frac{1-\exp(2c(0)(T-t)}{2c(0)}\right\}} \; x(t) \tag{5.16}$$

By definition $\bar{u}(t)$ is superior to $u^*(t)$. In general, $u^*(t)$ is superior to $u^{**}(t)$.

The system in Eq. (5.1) was simulated on a digital computer with the control schemes $u^*(t)$, $u^{**}(t)$ (with $c(0)=0$) and $\bar{u}(t)$. The discretization interval used was $\Delta=0.01$ sec. In Figure 1, the graph of the approximate optimal control $u^*(t)$, the corresponding trajectory $x(t)$, the conditional mean $c(t)$ and the conditional variance $p(t)$ are plotted versus time t. In Table 1, the value of the criterion function $J(u)$ is plotted for each of the control schemes $u^*(t)$, $u^{**}(t)$ and $\bar{u}(t)$, for different runs.

Table 1. The Value of the Criterion Function with Different Control Schemes for Different Runs.

Expt. No.	1	2	3	4	5	6	7	8	Average
Control $u^*(t)$	1.05	.287	.307	.648	.823	.851	.513	1.49	.746
Control $u^{**}(t)$	1.12	.379	.411	.652	.827	.873	.553	1.47	.786
Control $\bar{u}(t)$	.976	.0964	.139	.608	.791	.793	.451	1.50	.668

We note that the random numbers used to represent the noise for all the runs in the same column in Table 1 are identical.

From Table 1, we see that the control scheme of the paper $u^*(t)$ is uniformly superior to the adhoc scheme $u^{**}(t)$ for all the runs. The third row indicates the cost we would have obtained if c^* was known. As expected, this control leads to the least cost.

VI. DISCUSSION

The paper highlights the role of optimal estimation in the optimal control problems involving unknown parameters. The properties of the optimal estimates of $\underline{c}^*$ enables us to replace the unknown term $\underline{c}^*$ in the system equation (2.1) by its estimate $\underline{c}(t)$ and the innovation process $d\underline{y}(t)$. This is the key step of the paper. From this point on, the solution of the problem by the use of the principle of optimality is straight forward since there are no unknown parameters in the equivalent system equations.

The important feature of the control scheme is the absence of the need to store the entire measurement history $\underline{x}(t)$. At any instant t, the relevant information about $\{\underline{x}(v),\ 0 \leq v \leq t\}$ is stored in the parameters $\underline{c}(t)$ and $P(t)$, the conditional mean and variance of the unknown parameters, respectively. These parameters can be recursively computed. Thus, the

optimal feedback control at any instant t is a nonlinear function of t, $\underline{x}(t)$, $\underline{c}(t)$ and $P(t)$.

The exact optimal feedback control function involves the solution of a PDE for the optimal cost function $V(t, \underline{x}, \underline{c}, P)$, which is not easy to obtain. However, one can get approximate solutions by a power series in the variables P_{ij}, x_i. The larger the number of terms in the approximating function, the closer is the approximation to the true value. We can implement the control scheme in real time since the control at any instant does not involve any information that is not available at that instant. However, the gains like $g_1(t, \underline{c})$, $g_2(t, \underline{c})$ in the control scheme $\underline{u}^*(t)$ are obtained by integrating an ordinary differential equation from time t_f to the current time t. Thus, the amount of computation needed to implement the scheme is considerable.

REFERENCES

(1) K. Ito, Lectures on Stochastic Processes, Lecture notes, Tata of Fundamental Research, Bombay, 1961.

(2) T. Kailath and P. Frost, "Mathematical Modeling of Stochastic Processes" Sympoisum on "Stochastic Problems in Control," 1968, Published by American Society of Mechanical Engineers.

(3) J. J. Florentin, "Optimal Control of Continuous-Time, Markov, Stochastic Systems," J. Elec. and Control, 10, 1961, pp. 437-488.

(4) E. Wong and M. Zakai, "On the Relation Between Ordinary and Stochastic Differential Equations," Int. J. Eng. Sci., Vol. 3, pp. 213-229, 1965.

(5) R. E. Kalman and R. E. Bucy, "New Results in Linear Filtering and Prediction Theory," J. Basic. Eng. A.S.M.E., Series D, pp. 95-108, 1961.

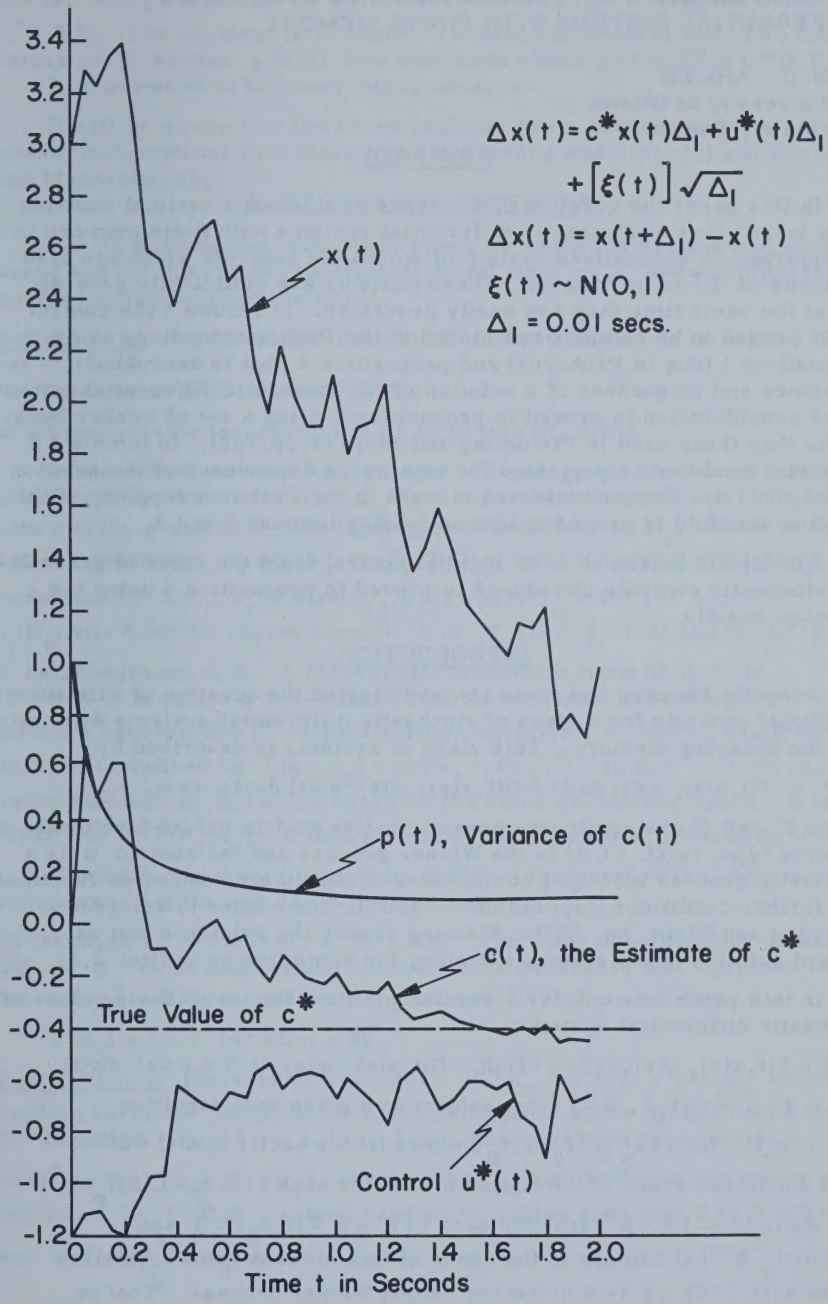


FIGURE I. GRAPH OF $x(t)$, $u^*(t)$, $c(t)$, $p(t)$ VERSUS TIME t.

OPTIMAL CONTROL OF A LARGE CLASS OF STOCHASTIC DIFFERENTIAL SYSTEMS WITH FINITE MEMORY

N.U. AHMED
University of Ottawa
Ottawa, Canada.

ABSTRACT

In this paper the question of existence of stochastic optimal controls for a large class of stochastic differential systems with finite memory is considered. A generalized system of stochastic controls which are also elements of L^∞ is introduced. These controls are sufficiently general and at the same time they are easily generated. In lemma 1 the control set is proved to be compact and closed in the Prohorov topology using proposition 1 (due to Prohorov) and proposition 2 (due to Skorokhod). Existence and uniqueness of a solution of the stochastic differential system under consideration is proved in proposition 3 using a set of weaker hypothesis than those used in Flemming and Nisio (1, pp. 782). In lemma 2 a sufficient condition is presented for continuous dependence of the solution on the control. Compactness and closure in the Prohorov topology of the solution manifold is proved in lemma 3 using lemmas 1 and 2.

Finally the existence of an optimal control from the class of generalized stochastic controls introduced is proved in proposition 4 using the previous results.

INTRODUCTION

Recently Fleming and Nisio (1) investigated the question of existence of optimal controls for a class of stochastic differential systems with infinite but decaying memory. This class of systems is described by

$$dx(t) = F(t,x(s);\ s \le t)\, du(t) + G(t,x(s),\ w(s);\ s \le t)\, dw(t),\quad t \ge 0,$$

where F and G are continuous nonanticipative matrix valued functionals of volterra type, $(w(t),\ t \in R)$ is the Wiener process and the control u is a stochastic process belonging to the class of absolutely continuous functions. For further conditions imposed on F and G the reader is referred to (Fleming and Nisio, pp. 782). Fleming proves the existence and uniqueness of solution of the system 1.1 along the same line as in (Ito, 2,3).

In this paper we consider a similar problem for the following class of stochastic differential systems:

$$S\ \begin{cases} dx(t) = F(t,x(s),\ u(s):\ t-T \le s \le t)\, dt + G(t,x(s),\ w(s);\ t-T \le s \le t)\ dw(t), \\ \qquad t \in I = [t_0\ t_1],\ 0 \le T < \infty;\ \text{subject to a given past condition} \\ x(t) = x_{\_}(t) \ \text{ for } t \in I_p \triangleq [t_0 - T, t_0] \end{cases}$$

where (i) for each $t \in I$ $x(t) \in E^n$ (real Euclidean space of dimension n) (ii) for each $t \in I$, $x \in C(I \cup I_p, E^n)$ and $u \in L^\infty(I, E^r)$ $F \in E^n$ (iii) for each $t \in I$, $x \in C(I \cup I_p, E^n)$ and $w \in C(I \cup I_p, E^m)$ G belongs to the space of (n x m) real valued matrices where $w(t)$, $t \in I_p \cup I$ is a m-vector valued Wiener process. That is G is a matrix valued functional. Let U be any compact subset of E^r

and let $K = \{ u \in L^{\infty}(I, E^r)$: $u(t) \in U$ for each $t \in I \}$ be a compact subset of $L^{\infty}(I, E^r)$. The problem is to choose a stochastic process $u(t)$, $t \in I$ taking values in K so that $\phi(x,u)$ is a minimum where $\phi: C(I, E^n) \times L^{\infty}(I, E^r) \rightarrow R$ is assumed to be lower semicontinuous.

It will be shown that the above problem has a solution under a set of weaker assumptions than those used in Fleming and Nisio (1) and Becker and Mandrekar (6).

2. STOCHASTIC CONTROLS

Let (Ω, B, P) be the probability space on which the system (x, w) is based and let $B(K)$ be the topological Borel field on K and $(K, B(K))$ the corresponding measurable space.

Definition 1: A generalized system of stochastic controls is defined by the set

$$M = \{u;\ u:\ (\Omega, B, P) \rightarrow (K, B(K))\}$$

of all measurable transforms of (Ω, B, P) into $(K, B(K))$.

In the sequel we need the concepts of topology for stochastic processes introduced by Prohorov. Let Σ_ρ be a separable complete metric space with metric ρ and $B(\Sigma_\rho)$ the topological Borel field on Σ_ρ. Given two probability measures μ_1 and μ_2 on $(\Sigma_\rho, B(\Sigma_\rho))$ the Prohorov metric $L(\mu_1, \mu_2)$ is defined by $L(\mu_1, \mu_2) = \max(\delta_{12}, \delta_{21})$ where $\delta_{ij} = \inf \{\delta:\ \mu_i(A) \leq \mu_j(N_\delta(A)) + \delta$ for all closed subsets A of $\Sigma_\rho\}$ $(i \neq j = 1, 2)$ and $N_\delta(A)$ is the δ neighbourhood of A. A measurable transform from (Ω, B, P) to $(\Sigma_\rho, B(\Sigma_\rho))$ is called a Σ_ρ - valued random variable. The Prohorov distance between two Σ_ρ -valued random variables x_1 and x_2 based on (Ω, B, P) is defined by $L(x_1, x_2) = L(Px_1^{-1}, Px_2^{-1})$. In fact it is not necessary that both x_1 and x_2 be based on the same probability space. It is clear that if $L(x_1, x_2) = 0$ then x_1 and x_2 have the same probability law.

The following theorems will be useful in the proof of the closure and compactness of the set M in the L-topology.

Proposition 1: (Prohorov, 4, th 1.12). For a set of probability measures $\mathcal{M}$ on Σ_ρ to be sequentially compact it is necessary and sufficient that for every $\epsilon > 0$ there exists a compact set $K_\epsilon \subset \Sigma_\rho$ such that

$$\mu(K_\epsilon) > 1 - \epsilon \text{ for all } \mu \in \mathcal{M}.$$

Proposition 2: (Skorokhod 5 pp. 281) If $\{u_n, n = 1, 2, \ldots\}$ are Σ_ρ valued random variables based on the same probability space (Ω, B, P) such that the sequence $\{Pu_n^{-1}\ n = 1, 2, \ldots\}$ is L-compact then there exists a subsequence $\{u_{n_k}\}$ of $\{u_n\}$ with L-limit u_0 and further there exists a sequence $\{u'_{n_k}\}$ of Σ_ρ - valued random variables such that $L(u_{n_k}, u'_{n_k}) = 0$ and $P\{\rho(u'_{n_k}, u_0) \rightarrow 0\} = 1$.

We are now prepared to prove the following lemma required in the sequel.

Lemma 1: The set M (Definition 1) is closed and compact in the L-topology.

Proof: Let $\{u_n\} \in M$ and let $u_n \to u_o$ P-a.e. Then since $\overline{\lim}\ u_n$ and $\underline{\lim}\ u_n$ are B measurable u_o is also B measurable. Further since U is a compact subset of E^r and K is assumed to be a compact subset of $L^\infty(I, E^r)$ K is closed and compact. Hence it follows that $u_o \in K$ P-a.e. Therefore $u_o \in M$ proving the closure of M.

Since K is compact and closed it follows from theorem 1 that $\{Pu_n^{-1}\}$ is L-compact for every sequence $\{u_n\} \in M$. Hence by theorem 2 we obtain a measurable transform u_o and a subsequence $\{u_{n_k}\} \subset \{u_n\}$ such that $P\{\sigma \in \Omega : \|u_{n_k} - u_o\| \to 0\} = 1$. Thus M is L-compact. This completes the proof of the lemma.

Some Notations: (i) No distinction will be made in the notation for the norm of a matrix and a vector. If z is a matrix then its norm $|z| \triangleq \sqrt{\sum_{ij} (z_{ij})^2}$ and if it is a vector then its norm again denoted by $|z|$ is given by $\sqrt{\sum_i (z_i)^2}$. The notation $\|\cdot\|$ will be preserved for the norm of a function belonging to a normed space.

(ii) Given a stochastic process $\{x(t), t \in I\}$ $B_{rs}(x)$ denotes the least Borel algebra generated by $\{x(t), r \le t \le s\}$. The m-dimensional wiener process is denoted by $w(t) = (w_1(t), \ldots w_m(t))$, $t \in I$ and we normalize it by $w(t_o) = 0$. $B_{rs}(dw)$ denotes the least Borel algebra generated by $\{w(t) - w(\tau), r \le \tau \le t \le s\}$. The least Borel algebra that contains $B_1, B_2 \ldots B_n$ is denoted by $B_1 \vee B_2 \vee \ldots B_n$.

Let $\{x_\_(t), t \in I_p\}$ be a given stochastic process with continuous paths in E^n and let $B_\_$ denote the corresponding Borel algebra.

Definition 2: A generalized stochastic control $u \in M$ is called an admissible control or more precisely the triple $(x_\_, u, w)$ is called an admissible system if $B_\_ \vee B_{t_o,t}(u) \vee B_{t_o - T, t}(w)$ is independent of $B_{t,t_1}(dw)$ for every $t \in I$. In fact the set M itself can be regarded as the set of all admissible generalized stochastic controls.

A stochastic process $\{x(t), t \in I_p \cup I\}$ with continuous paths is called a solution of the system S(for an admissible system $(x_\_, u, w)$ with the past condition $x_\_$ if

$x(t) = x_\_(t)$ for $t \in I_p$ with probability one and $B_{t_o - T, t}(x) \vee B_{t_o, t}(u) \vee B_{t_o - T, t}(w)$ is independent of $B_{t,t_1}(dw)$ for every $t \in I$;

and if

$$x(t) = x(to) + \int_{to}^{t} F(\tau, x(s), u(s);\ \tau - T \le s \le \tau)d\tau + \int_{to}^{t} G(\tau, x(s), w(s);\ \tau - T \le s \le \tau)\, dw(\tau)\ t \in I .$$

The first integral on the right is regarded as ordinary Lebesgue integral

and the second as the I to integral (3, pp 14).

3. EXISTENCE AND UNIQUENESS OF SOLUTION

Before considering the question of existence of stochastic optimal control it will be necessary to investigate into the question of existence and uniqueness of solution of the system S. For this purpose we make the following assumptions:

A_1: There exist two real valued functions H_1 and $H_2 \in L^2(IXI')$ (where $I' \triangleq I \cup I_p$ with Lebesque measure $\mu(I) < \infty$) so that for almost all $t \in I$

$$\left| F(t,x(s),\ u(s);\ t-T \le s \le t) - F(t,y(s),\ u(s);\ t-T \le s \le t) \right|$$
$$\le \int_{t-T}^{t} \left| H_1(t,s) \right| \left| x(s) - y(s) \right| ds$$

$$\left| G(t,x(s),\ w(s);\ t-T \le s \le t) - G(t,y(s),\ w(s);\ t-T \le s \le t) \right|$$
$$\le \int_{t-T}^{t} \left| H_2(t,s) \right| \left| x(s) - y(s) \right| ds$$

for all $u \in K$ and $w \in C(I', E^m)$.

A_2: There exist two real valued functions α_1 and $\alpha_2 \in L^2(I)$ and a real valued function $\beta_2 \in L^2(IXI')$ and an integer $q \ge 1$ so that for almost all $t \in I$

$$\left| F(t,o,\ u(s);\ t-T \le s \le t) \right| \le \left| \alpha_1(t) \right| \text{ for all } u \in K$$

and

$$\left| G(t,o,\ w(s);\ t-T \le s \le t) \right| \le \left| \alpha_2(t) \right| + \int_{t-T}^{t} \left| \beta_2(t,s) \right| \left| w(s) \right|^q ds.$$

A_3: $\sup_{t \in I_p} \mathcal{E} \left| x_-(t) \right|^2 \le C^2 < \infty$.

The following proposition will be found useful in the sequel.

Proposition 3: Under the assumptions A_1, A_2 and A_3; and for each admissible generalized stochastic control $u \in M$ there exists one and only one solution of the system S with locally bounded second moments and this solution satisfies the property $B_{to-T,t}(x) \subset B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w)$.

Proof: It is enough to solve the stochastic integral equation

$$x(t) = x_-(to) + \int_{to}^{t} F(\tau, x(s),\ u(s);\ \tau - T \le s \le \tau) d\tau + \int_{to}^{t} G(\tau\ x(s),\ w(s);$$
$$\tau - T \le s \le \tau)\ dw(\tau) \quad t \in I.$$

To solve this equation we will use the well known successive approximation method (Ito and Nisio pp 41, Fleming and Nisio, pp 783). Let $\{ x_o(t),\ t \in I' \}$ be any continuous stochastic process satisfying only the properties: (i) $\sup_{t \in I'} \mathcal{E} \left| x_o(t) \right|^2 \le a_o^2 < \infty$, (ii) $x_o(t) = x_-(t)$ for $t \in I_p$ with probability one and (iii) $B_{to-T,t}(x_o)$ is independent of $B_{t,t_1}(dw)$ with $B_{to-T,t}(x_o) \subset B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w)$.

Let us also define the sequence of random processes $\{ x_n(t), t \in I' \}$, $n = 1, 2, \ldots$ by $x_n(t) = x_-(to) + \int_{to}^{t} F(\tau, x_{n-1}(s),\ u(s);\ \tau - T \le s \le \tau) d\tau +$

$$+ \int_{to}^{t} G(\tau, x_{n-1}(s), w(s); \tau - T \le s \le \tau)\, dw(\tau) \qquad 3.2$$

so that for each n, $x_n(t) = x_{\_}(t)$ for $t \in I_p$ with probability one.

It follows from the definition of I to integral (Ito, pp 14) and 3.2 that

$$\mathcal{E}|x_1(t)|^2 \le 3\mathcal{E}|x_{\_}(to)|^2 + 3(t-to)\int_{to}^{t} \mathcal{E}|F(\tau, x_o(s), u(s); \tau - T \le s \le \tau)|^2 d\tau$$

$$+ 3m \int_{to}^{t} \mathcal{E}|G(\tau, x_o(s), w(s); \tau - T \le s \le \tau)|^2 d\tau \qquad 3.3$$

By assumptions A_1 and A_2 we have

$$\int_I \mathcal{E}|F(\tau, x_o(s), u(s); \tau - T \le s \le \tau)|^2 d\tau \le 2(\|a_1\|^2 + T a_o^2 \|H_1\|^2) \qquad 3.4$$

and $\int_I \mathcal{E}|G(\tau, x_o(s), w(s); \tau - T \le s \le \tau)|^2 d\tau \le 2(\|a_2\|^2 + 2a_o^2 \|H_2\|^2 +$

$$\frac{2^{q+1}}{\sqrt{\pi}}\, m^q\, \Gamma(q+\tfrac{1}{2})\, \|\beta_2\|^2 \cdot \int_{to-T}^{t_1} |s - to|^q\, ds) \qquad 3.5$$

where

$$\|a_i\|^2 = \int_I |a_i(t)|^2\, dt. \quad \|H_i\|^2 = \int_{I \times I'} \int |H_i(t,\tau)|^2 d\tau\, dt \quad i = 1, 2 \text{ and}$$

$$\|\beta_2\|^2 = \int_{I \times I'} \int |\beta_2(t,\tau)|^2\, d\tau dt.$$

From 3.3, 3.4 and 3.5 and the assumption A_3 it follows that there exists a positive real number a_1^2 such that $\sup_{t \in I'} \mathcal{E}|x_1(t)|^2 \le a_1^2 < \infty$.

Defining $\rho_n(t)$, $n = 1, 2, \ldots$ by

$\rho_n^2(t) = \sup_{to \le s \le t} \mathcal{E}|x_n(s) - x_{n-1}(s)|^2$ it follows from the above discussion and the obvious inequality

$$\mathcal{E}|x_1(t) - x_o(t)|^2 \le 2\mathcal{E}|x_1(t)|^2 + 2\mathcal{E}|x_o(t)|^2 \le 2(a_1^2 + a_o^2) \triangleq a_2^2$$

for all $t \in I'$ that

$$\rho_1^2(t) \le a_2^2 < \infty \quad \text{for all } t \in I. \qquad 3.6$$

Using Ito integral rule once again and the abreviations

$F(t,x,u)$ for $F(t, x(s), u(s); t - T \le s \le t)$ and $G(t,x,w)$ for $G(t, x(s), w(s); t - T \le s \le t)$ we obtain $\rho_{n+1}^2(t)$

$$\le 2(t - to + m) \int_{to}^{t} d\tau\, \mathcal{E}(|F(\tau, x_n, u) - F(\tau, x_{n-1}, u)|^2$$

$$+ |G(\tau, x_n, w) - G(\tau, x_{n-1}, w)|^2) \qquad 3.7$$

for all $t \in I$.

By use of the assumption A_1 and the fact that $\sup_{t \in I_p} \mathcal{E}|x_n(t) - x_{n-1}(t)|^2 = 0$ $n = 1, 2, \ldots$ we obtain from 3.7 that

$$\rho_{n+1}^2(t) \le 2T(t - to + m) \int_{to}^{t} B^2(\tau)\, \rho_n^2(\tau)\, d\tau, \qquad 3.8$$

where $B^2(\tau) \triangleq \int_{\tau - T}^{\tau} (|H_1(\tau, s)|^2 + |H_2(\tau, s)|^2)\, ds$.

Since H_1 and $H_2 \in L^2(IxI')$ (by hypothesis) $B^2 \in L'(I)$ and therefore $\rho_2^2(t) \le 2T(t\text{-}to+m)\, a_2^2\, h(t)$ where $h(t) \triangleq \int_{to}^{t} B^2(\tau)\, d\tau$ is obviously absolutely continuous on I. Consequently it follows from 3.8 that

$$\rho_{n+1}^2(t) \le a_2^2\, (2T(t\text{-}to+m))^n h^n(t) / n!\ , \quad n = 1,2,\ldots \text{ for all } t \in I. \qquad 3.9$$

Combining Tschebycheff's inequality and the inequality 3.9 it is easily verified that for every $\epsilon > 0$

$$P\{ \sup_{to \le s \le t} |x_{n+1}(s) - x_n(s)| > \epsilon \} \le \frac{1}{\epsilon^2}\, \rho_{n+1}^2(t) \qquad 3.10$$

and for $\epsilon = 2^{-n}$ we have

$$P\{ \sup_{to \le s \le t} |x_{n+1}(s) - x_n(s)| > 2^{-n} \} \le a_2^2\, \frac{(8T\,(t\text{-}to+m))^n h^n(t)}{n!}$$

Hence by Borel-Cantelli's lemma

$P\{ \sup_{to \le s \le t} |x_{n+1}(s) - x_n(s)| \le 2^{-n}$ for large $n \} = 1$ and therefore

$$P\{ \sum_{n=o}^{\infty} \sup_{to \le s \le t} |x_{n+1}(s) - x_n(s)| < \infty \} = 1.$$

This shows that with probability one

$$x_n(t) = x_o(t) + \sum_{i=1}^{n} (x_i(t) - x_{i-1}(t)), \quad n = 1,2,\ldots \qquad 3.11$$

is convergent uniformly on every bounded interval $I \subset [\, to, \infty\,)$. Therefore the limiting process $\{x(t),\ t \in I\}$ given by $x(t) \equiv \lim_n x_n(t)$ exists and is continuous in t with probability one. It is clear that

$$P\{ \sup_{t \in I} |x_n(t) - x(t)| \to 0 \} = 1. \qquad 3.12$$

By use of the hypothesis A_1 it is easily verified that with probability one

$$\int_{to}^{t} F(\tau, x_n(s), w(s);\ \tau\text{-}T \le s \le \tau)\, d\tau \to \int_{to}^{t} F(\tau, x(s), u(s);\ \tau\text{-}T \le s \le \tau)\, d\tau$$

and

$$\int_{to}^{t} G(\tau, x_n(s), w(s);\ \tau\text{-}T \le s \le \tau)\, dw(\tau) \to \int_{to}^{t} G(\tau, x(s), w(s);\ \tau\text{-}T \le s \le \tau)\, dw(\tau)$$

uniformly on I. Therefore we conclude that x(ie. the process $\{x(t), t \in I\}$ is a solution of the system S. That this solution has bounded second moment is proved as follows:

Let us define $d_n^2(t) \triangleq \sup_{to \le s \le t} E|x_n(s)|^2 \qquad n = 0,1,2,\ldots$

By use of the assumption A_2 it is readily verified that

$$d_n^2(t) \le K^2(t) + 6(t\text{-}to+m)\, T \int_{to}^{t} B^2(\tau)\, d_{n-1}^2(\tau)\, d\tau\ , \quad n = 1,2,\ldots \qquad 3.13$$

$t \in I$ where $K^2(t)$ defined as

$$K^2(t) \triangleq 3C^2 + 6(t\text{-}to+m) \int_{to}^{t} E(\,|F(\tau,0,u)|^2 + |G(\tau,0,w)|^2)\, d\tau$$

is finite for all $t \in I$. Further by successively substituting 3.13 into itself one is convinced that

$$d_n^2(t) \le K^2(t) \Big(\sum_{\gamma=o}^{n} (6T\,(t\text{-}to+m)\, h(t))^{\gamma} / \gamma\,! \Big) + \frac{6T(t\text{-}to+m)\, h(t)\, d_o^2(t)}{n!}\ .$$

Hence

$$\sup_{to \le s \le t} \mathcal{E}|x(s)|^2 \le 2K^2(t)\, e^{6T(t-to+m)\, h(t)} \qquad 3.14$$

for all $t \in I$. Thus the solution x has bounded second moment on any bounded interval $I \subset [to, \infty)$ and is independent of the choice of the process $x_o(t)$, $t \in I$.

The uniqueness of the solution is proved as follows. Let x be the solution obtained above and let y be any other solution with bounded second moment so that $y(t) = x(t)$ for $t \in I$ with probability one.

Defining $\rho^2(t) \triangleq \sup_{to \le s \le t} \mathcal{E}|x(s) - y(s)|^2$, $t \in I$ 3.15

and using the same procedure as followed in constructing the inequality 3.8 we obtain

$$\rho^2(t) \le 2T(t - to + m) \int_{to}^{t} B^2(\tau)\, \rho^2(\tau)\, d\tau, \quad t \in I. \qquad 3.16$$

By use of Gronwalls Lemma it follows from 3.16 that $\rho^2(t) \equiv 0$ for all $t \in I$ and for all finite intervals $I \subset [to, \infty)$. Thus

$$P\{x(t) - y(t) = 0\ t \in I\} = 1.$$

We conclude the proof of this proposition by showing the validity of the inclusion relation for the Borel algebras.

From 3.2 and the initial condition $B_{to-T,t}(x_o) \subset B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w)$ it follows that for each n

$$B_{to-T,t}(x_n) \subset B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w).$$

since for each fixed t, $x_n(t) \to x(t)$ with probability one we have $\overline{\lim} x_n(t) = \underline{\lim} x_n(t) = x(t)$ with probability one. Since $\overline{\lim} x_n(t)$ and $\underline{\lim} x_n(t)$ are $B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w)$ measurable the limit $x(t)$ is $B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w)$ measurable. This is true for each $t \in I$ so we have

$$B_{to-T,t}(x) \subset B_- \vee B_{to,t}(u) \vee B_{to-T,t}(w). \qquad 3.17$$

Remark: It may be interesting to note that the hypothesis on G under the assumption A_2 can be weakened to include nonlinearities of the form $O(e^{a|w|^2})$.

The following lemma will be found useful in the sequel.

Lemma 2: If for each fixed $x \in C(I', E^n)$ F is a strongly continuous operator (in u) from $L^\infty(I, E^r)$ to $L^2(I', E^n)$ uniformly on $K \subset L^\infty(I, E^r)$ and further if F and G satisfy the conditions A_1, A_2 and A_3 then the solution x of the system S is continuously dependent on the control u.

Proof: Let $\{u_n\}$ by any sequence in K converging strongly to an element $u^* \in K$ and let $\{x_n\}$ and x^* be the unique solutions (in the sense of the previous theorem) of the system S corresponding to $\{u_n\}$ and u^* respectively. We will obtain the proof of the lemma by demonstrating that $x_n \to x^*$ p-a.e uniformly on I. For this purpose let us define $\delta_n^2(t) \triangleq \sup_{to \le s \le t} \mathcal{E}|x^*(s) - x_n(s)|^2$ for $t \in I$ and $n = 1, 2, \ldots$.

Clearly from the above discussion and the integral equation 3.1 it follows that

$$\mathcal{E}\,|x^*(s) - x_n(s)|^2 \le 3(s\text{-}to)\,\mathcal{E}\int_{to}^{s} |F(\tau, x^*(\theta), u^*(\theta); \tau\text{-}T \le \theta \le \tau) - F(\tau, x^*(\theta), u_n(\theta);$$
$$\tau\text{-}T \le \theta \le \tau)|^2\, d\tau$$
$$+\ 3(s\text{-}to)\,\mathcal{E}\int_{to}^{s} |F(\tau, x^*(\theta), u_n(\theta); \tau - T \le \theta \le \tau) - F(\tau, x_n(\theta), u_n(\theta);$$
$$\tau\text{-}T \le \theta \le \tau)|^2\, d\tau$$
$$+\ 3m \int_{to}^{s} \mathcal{E}\,|G(\tau, x^*(\theta), w(\theta); \tau\text{-}T \le \theta \le \theta \le \tau) - G(\tau, x_n(\theta), w(\theta);$$
$$\tau\text{-}T \le \theta \le \tau)|^2\, d\tau \qquad 3.18$$

Therefore by assumption A_1 we have for $t \in I$

$$\delta_n^2(t) \le 3(t\text{-}to)\,\mathcal{E}\int_{to}^{t} |F(\tau, x^*(\theta), u^*(\theta); \tau\text{-}T \le \theta \le \tau) - F(\tau, x^*(\theta), u_n(\theta);$$
$$\tau\text{-}T \le \theta \le \tau)|^2\, d\tau$$
$$+\ 3T(t - to + m) \int_{to}^{t} B^2(\tau)\, \delta_n^2(\tau)\, d\tau, \qquad 3.19$$

where $B^2(t)$ is given by the expression following the inequality 3.8. For the sake of abreviation let us define

$$f_n(t) \triangleq \int_{to}^{t} |F(\tau, x^*(\theta), u^*(\theta); \tau - T \le \theta \le \tau) - F(\tau, x^*(\theta), u_n(\theta);$$
$$\tau - T \le \theta \le \tau)|^2\, d\tau$$

Since $u_n, u^* \in K$ and x^* is a solution of the system S with bounded second moment it is easily verified that $f_n(t) < \infty$ for all $t \in I$ and for all n. It is clear that $f_n(t') \le f_n(t'')$ for $t' \le t''$ for all n. Further since F is a strongly continuous operator from L^∞ to L^2 uniformly on K and since x^* is continuous with probability one on I (by proposition 3), we have that $f_n(t) \to 0$ p.a.e. uniformly on I. Further by Fatou's lemma (10) we have for each $t \in I$

$$\mathcal{E}\,\underline{\lim}\, f_n(t) \le \underline{\lim}\, \mathcal{E} f_n(t) \le \lim \mathcal{E} f_n(t) \le \overline{\lim}\, \mathcal{E} f_n(t) \le \mathcal{E}\,\overline{\lim}\, f_n(t).$$

Therefore $\lim \mathcal{E} f_n(t) = 0$ for all $t \in I$.

Again it follows from Gronwall's lemma applied to 3.19 that

$$\delta_n^2(t) \le 3(t_1 - t_0)\,\mathcal{E} f_n(t_1).\ e^{3T(t_1 - t_0 + m) \int_{to}^{t} B^2(\tau)\, d\tau} \qquad 3.20$$

for all $t \in I$.
Since $3(t_1 - t_0)\, e^{3T(t_1 - t_0 + m \int_{to}^{t} B^2(\tau) d\tau}$ is finite for each $t \in I$, we have $\lim \delta_n^2(t) = 0$ for each $t \in I$. Therefore $x_n \to x^*$ p-a.e uniformly on I. This completes the proof of the lemma.

4. EXISTENCE OF STOCHASTIC OPTIMAL CONTROL

We are now prepared to prove the existence of an optimal control for the system S.

Let us define a set $\mathcal{X}$ for the wiener process $\{ w(t),\ t \in I' \}$ by

$$\mathcal{X} = \{ x \in C(I', E^n) :\ x(t) = x(to) + \int_{to}^{t} F(\tau, x(s), u(s); \tau \le s \le \tau)\, d\tau$$
$$+ \int_{to}^{t} G(\tau, x(s), w(s); \tau - T \le s \le \tau) dw(\tau);\ x(t) = x_-(t) \text{ for } t \in I_p \text{ and } u \in M \} \qquad 4.1$$

Lemma 3: The set $\mathcal{X}$ is L-compact and L-closed .

Proof: Since M is L-compact and L-closed and $\mathcal{X}$ is a continuous image of M under x by lemma 2, we conclude that $\mathcal{X}$ is L-compact and L-closed.

The optimal control problem stated in the introduction is solved in the following.

Proposition 4: If φ is a real valued non-negative lower-semicontinuous cost functional defined on $\mathcal{X} \times M$ then there exists a control $u^* \in M$ such that the expected value $\mathcal{E}\varphi(x^*, u^*)$ is the minimum.

Proof: Follows from L-compactness of $\mathcal{X} \times M$ and lower semicontinuity of $\mathcal{E}\phi$ (Ahmed 12).

In common practice the cost functional φ is given as an integral of the form

$$\varphi(x,u) = \int_I F^o(t, x(t), u(t))\ dt$$

where F^o is any suitable non negative real valued continuous function on $I \times E^n \times E^r$. Clearly the functional φ is lower - semicontinuous. In fact this is true even when F^o is merely lower-semicontinuous in x and u for almost all $t \in I$ and measurable and integrable in t on I for every fixed $(x,u) \in E^n \times E^r$.

An interesting example of a linear stochastic heriditary system is presented below:

$$dx(t) = (K_o(t) + \int_{t=T}^{t} K_1(t,s)\, x(s)ds + K_2(t)\, u(t))dt + K_3(t)\, dw(t),\ t \in I.$$

where K_o, K_1, K_2, K_3, x, u are of appropriate dimension and that the elements of K_o, K_1, K_2 and K_3 are locally Lebesgue square integrable. The results of this paper apply to this system in a straight-forward manner. If $K_3 = 0$ then the above system reduces to a deterministic one, further it can be shown that if the L^2 norm of K_3 is infinitesimaly small then the solution of the stochastic system differs from that of the deterministic one only infinitesimaly.

Remarks:

(i) The criterion of compactness of a bounded subset $K \subset L^\infty$ is given in Dunford and Schwartz (11). It is felt that this class of controls can be relatively easily generated than those suggested by Fleming and Nisio (1) or Becker and Mandrekar (6).

(ii) The problem of considering feedback control $u(t) = v(t, x(s);\ s \le t)$ based on the state is more interesting and is not known to be satisfactorily resolved.

Conclusion: In this paper we have considered the question of existence of stochastic optimal controls for a large class of stochastic heriditary differential equations employing sufficiently weaker hypothesis than those used by Fleming and Nisio (1) or Becker and Mandrekar (6). The class of controls chosen is also relatively easily implement able.

REFERENCES

(1) W.H. Fleming and M. Nisio; "On the existence of Optimal Stochastic Controls", Journal of Mathematics and Mechanics, Vol. 15, No. 5, 1966, pp. 777-794.

(2) K. Ito and M. Nisio, "On Stationary Solutions of a Stochastic Differential Equation", J. Math., Kyoto Univ., 4-1 (1964), pp. 1-75.

(3) K. Ito, "On Stochastic Differential Equations", Mem. Am. Math. Soc., No. 4, 1951.

(4) Yu, V. Prochorov, "Convergence of Random Processes and Limit Theorems in Probability Theory", Theory of Probability and Its Applications, 1, 1956, pp. 157-214.

(5) A.V. Skorokhod, "Limit Theorems for Stochastic Processes", Theory of Probability and its Application, 1, 1956, pp. 261-289.

(6) H. Becker and V. Mandrekar, "On the Existence of Optimal Random Controls", MRC Technical Summary Report No. 895, June 1968, pp. 1-23.

(7) H.J. Kushner, "Stochastic Stability and Control", Chapter 4, Acad. Press N.Y., 1967.

(8) N.U. Ahmed, "A Class of Stochastic Nonlinear Integral Equations on L^P Spaces and its Application to Optimal Control", J. Information and Control, July 1969,

(9) N.U. Ahmed, "Closure and Completeness of Wiener's Orthogonal Set $\{G_n\}$ in the Class $L^2(\Omega, B, \mu)$ and its Application to Stochastic Heriditary Differential Systems", TR 69-3, June 1969, Univ. of Ottawa (submitted to J. Inf. and Control).

(10) S. K. Berberian, "Measure and Integration", MacMillan Company, N.Y., 1965, pp. 105.

(11) N. Dunford and J.T. Schwartz, "Linear Operators", Part I, Inter Sciences, N.Y., 1964, pp. 342.

(12) N.U. Ahmed "Optimal Control of Stochastic Differential Systems with Finite Memory" TR 69-5, U of Ottawa, Dept. of Elect. Engg, July 1969.

---

* This work was supported in part by the National Research Council of Canada under Grant No. A-7109.

AN OPTIMAL CONTROLLER FOR A PLANT WITH RANDOM PARAMETERS

H. JON THOMAS
Rochester Institute of Technology
Rochester, New York

JOHN B. LEWIS
Pennsylvania State University
University Park, Penna.

Abstract - A control law is developed for a linear, random-parameter system which minimizes the expectation of a given quadratic cost functional. The Maximum Principle is used and a Riccati-type differential equation is obtained which is used to give a feedback gain matrix to produce the control signal. A second-order example is given with several different weighting matrices in the cost functional.

I. INTRODUCTION

Optimization techniques for automatic control systems have been intensely studied for a wide variety of systems. In all cases, it is desired to minimize or maximize a given functional and to find the corresponding input to the given plant which yields the minimum value. Astrom (2) did early work in analyzing continuous systems with random parameters, and was interested in the input/output relationships. A number of authors (1,4, 6, 7, 9, 11, 14) have used dynamic programming to investigate discrete systems with random parameters. Drenick and Shaw (6), Sworder (13), and Wonham (15) have used dynamic programming to consider continuous random parameter systems; and Zadicario and Sivan (16) as well as Gururaja (10) have used the Maximum Principle to investigate continuous systems with constant unknown parameters.

This paper considers the problem of optimizing a continuous, random-parameter system, with a cost functional of the Bolza type. It is also desired to obtain the mean of the output of the system and to examine several suboptimal control policies to determine the deterioration in performance resulting from these suboptimal controls. Section II describes the plant model as well as the method of obtaining the output mean. Section III describes the derivation of the optimal control law, and Section IV presents an example of a system using several different cost functionals, and compares trajectories and the output mean resulting from these control policies.

II. MODEL DESCRIPTION AND INPUT/OUTPUT STATISTICS

Model Used in the Study

The plant model chosen is a single-input, single-output continuous time linear plant. It is of n^{th} order, and has at most n independent randomly varying parameters. The representation of this system is written as

$$y^{(n)}(t) + \sum_{i=1}^{n} a_i(t) y^{(n-i)}(t) = u(t) \qquad (1)$$

with $y(0)$, $y^{(1)}(0), \ldots y^{(n-1)}(0)$ given.

In Equation (1), the $a_i(t)$ are the random coefficient processes, and they are assumed to be statistically independent. It is assumed that the value of any of the random coefficient processes a_i does not depend on the value of any other process in the system. Thus $E(a_i(\tau) | y^{(k)}(\tau) = \alpha)$ is independent of k and α. This implies that the a_i processes are uncorrelated with y and its derivatives. Also, each of the $a_i(t)$ processes has known mean and autocorrelation. It will be seen that the computation of the output mean does not require the mutual independence assumed above, but this is required in the derivation of the optimal control law, and it is assumed true for that reason.

Output Statistics

The first order statistics yield a "mean trajectory" about which the actual trajectory can be expected to vary. To obtain the mean of the output, $\bar{y}(t)$, first take expectation of (1). If y and its derivatives are continuous in the mean-square sense, and if the $a_i(t)$ are uncorrelated with y and its derivatives, the result can be written (see Papoulis (12)):

$$\bar{y}^{(n)}(t) + \sum_{i=1}^{n} \bar{a}_i(t)\bar{y}^{(n-i)}(t) = u(t) \tag{2}$$

The initial conditions to be used in the solution of (2) are the known initial conditions of the plant, given in (1).

The equation defining the output autocorrelation can be found by writing (1) twice; once using t_1 as the independent variable, and again using t_2 as the independent variable. When these two equations are multiplied together, and the expectation of the product is taken, the result is

$$\sum_{i=0}^{n} R_i(t_1,t_2) \frac{\partial^{2i} R_y(t_1,t_2)}{\partial t_1^i \partial t_2^i} + \sum_{j=0}^{n} \sum_{\substack{k=0 \\ j\neq k}}^{n} \bar{a}_j \bar{a}_k \frac{\partial^{j+k} R_y(t_1,t_2)}{\partial t_1^j \partial t_2^k} = u(t_1)u(t_2) \tag{3}$$

where the $R_i(t_1,t_2)$ are the correlation functions of the coefficient processes, and the $\bar{a}_j(t_1)$ and $\bar{a}_k(t_2)$ are the mean of a_j and a_k respectively. The boundary conditions for (3) can be found from the solution to (2), and also from the solution to

$$R_o(T,t)R_y(T,t) + \bar{a}_o(T) \sum_{i=1}^{n} \bar{a}_i(t) \frac{\partial^i R_y(T,t)}{dt^i} = 0. \tag{4}$$

The first n-1 derivatives of $R_y(T,t)$ can be found from the solution of (2) and the initial conditions of (1). After finding the boundaries for $R_y(t_1,t_2)$, one can obtain a solution to (3) by using the corresponding difference equation and solving it on a digital computer (see, for example, Forsythe and Wasow (8)).

III. AN OPTIMAL CONTROL LAW

To obtain the optimal control law, the problem is expressed using state-variable notation. It is assumed that the plant

is completely state-observable, and that all state variables are available to the controller at all times during the fixed control interval. Then, since this particular problem is a regulator problem, the specific minimization problem is stated in the following form. Given a plant

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + \underline{b}u(t) \tag{5}$$

where $\underline{x}(0) = \underline{x}_o$ is the given initial vector state, $\underline{b}$ is the given drive vector, and $A(t)$ is the nxn plant matrix containing the randomly varying plant parameters; find a control law $u(\underline{x}(t),t)$ which minimizes the cost functional:

$$J = \frac{1}{2}E\left\{<\underline{x}(T),F\underline{x}(T)> + \int_0^T (<\underline{x}(t),Q\underline{x}(t)> + \lambda u^2(t))dt\right\} \tag{6}$$

In (6) F is a known, nxn constant matrix, Q is a known, possibly time-varying nxn matrix, and λ is a known constant. It is also known that F and Q(t) are positive definite matrices, and that $\lambda>0$. Because the random plant parameters are varying in time as well as in probability space, the i^{th} realization of the plant produces a particular matrix time function $A_i(t)$ which defines the plant over the control interval $t\epsilon[0,T]$ for that particular plant realization.

Description of the Optimization Method

It is assumed first that there exists a control policy $u^*(\underline{x}(t),t)$ for (5) which minimizes J in Equation (6). The actual value of the cost functional for the i^{th} plant realization driven by $u_i^*(t)$ is denoted by J_i^*. Now let $u_i^*(t)$ undergo a first-order variation, such that $u_i = u_i^* + \epsilon\eta(t)$ where $\eta(t)$ is a weak variation about $u_i^*(t)$ whose first n time derivatives are bounded. This weak variation about $u_i^*(t)$ will result in a trajectory $\underline{x}_i(t)$ which differs from the mean optimal trajectory by a function $\underline{\varphi}_i(t)$. As $\epsilon\to 0$, it can be shown that $\underline{\varphi}_i(t)$ is defined by

$$\dot{\underline{\varphi}}_i(t) = A_i(t)\underline{\varphi}_i(t) + \underline{b}\eta(t);\ \underline{\varphi}(0) = \underline{0} \tag{7}$$

From (7), since $\eta(t)$ is a continuous function everywhere in the interval $t\epsilon[0,T]$, and since the elements of $A_i(t)$ are finite for all i and t, $\underline{\varphi}_i(t)$ also is a continuous function in the interval $[0,T]$. The value of the cost functional for the i^{th} plant realization when driven by $u_i(t)$ is

$$J_i = \frac{1}{2}<\underline{x}_i(T),F\underline{x}_i(T)> + \frac{1}{2}\int_0^T (<\underline{x}_i(t),Q\underline{x}_i(t)> + \lambda u_i^2(t))dt \tag{8}$$

The value of the cost functional for the i^{th} plant realization when driven by the optimal control $u_i^*(t)$ is similar.

The Hamiltonian is defined as

$$H_i = \frac{1}{2}\left[<\underline{x}_i(t),Q\underline{x}_i(t)> + \lambda u_i^2(t)\right] + <\underline{p}_i(t),\dot{\underline{x}}_i(t)> \tag{9}$$

where $\underline{p}(t)$ is a vector process which has statistics to be determined from the minimization. If $u_i(t)$ is replaced by $u_i^*(t)$ and

$\underline{x}_i(t)$ is replaced by $\underline{x}_i^*(t)$ in (9), one has the defining equation for H_i^*, the Hamiltonian associated with the i^{th} plant realization when driven by the optimal control $u_i^*(t)$.

Now write the difference $J_i - J_i^*$ in terms of H_i and H_i^* as

$$J_i - J_i^* = \frac{1}{2}\left[<\underline{x}_i(T), F\underline{x}_i(T)> - <\underline{x}_i^*(T), F\underline{x}_i^*(T)>\right] + \int_0^T (H_i - H_i^*)dt$$

$$+ \int_0^T <\underline{p}_i(t), \dot{\underline{x}}_i^* - \dot{\underline{x}}_i>dt \qquad (10)$$

Since H_i is continuous in $\underline{x}$, u, and t, expand H_i in a Taylor's series about H_i^*. Then expand the weighting term on the terminal error in a Taylor's series, and substitute these expansions into (10) to give

$$J_i - J_i^* = \epsilon<F\underline{x}_i^*(T), \underline{\varphi}_i(T)> + \epsilon\int_0^T (<Q\underline{x}_i^*(t) + A_i^T\underline{p}_i(t), \underline{\varphi}_i(t)>$$

$$+ (\lambda u_i^*(t) + \underline{b}^T\underline{p}_i(t)\eta(t))dt - \epsilon\int_0^T <\underline{p}_i(t), \dot{\underline{\varphi}}_i(t)>dt \qquad (11)$$

Integrating the last term in (11) by parts, noting that $\underline{\varphi}_i(0) = \underline{0}$, collecting the remaining terms in (11), then taking expectation over the ensemble, and dividing by ϵ gives

$$\frac{E(J-J^*)}{\epsilon} = E\left\{<F\underline{x}^*(T) - \underline{p}(T), \underline{\varphi}(T)> + \int_0^T (<Q\underline{x}^*(t) + A^T\underline{p}(t) + \dot{\underline{p}}(t), \underline{\varphi}(t)>\right.$$

$$\left. + \eta(t)(\lambda u^*(t) + \underline{b}^T\underline{p}(t))dt\right\} \qquad (12)$$

Since the coefficient processes vary independently of all other processes, then they vary independently of the costate vector $\underline{p}$, and consequently,

$$E(A(t)\underline{p}(\tau)) = E(A(t))\ E(\underline{p}(\tau)) \qquad (13)$$

Since $\underline{\varphi}_i(0) = \underline{0}$ for all i, the solution to (8) is given over the ensemble by

$$\underline{\varphi}(t) = \int_0^t \Psi(\tau, t_o)\underline{b}\eta(\tau)d\tau \qquad (14)$$

where $\Psi(\tau, t_o)$ is a random transition matrix corresponding to A.

Since $\eta(t)$ is any arbitrary n-differentiable time function, then if c is any random variable which does not depend on $\eta(t)$,

$$E(c|\underline{\varphi}(t) = \underline{\xi}) = E(c). \qquad (15)$$

The reason is that for a given $\underline{\xi}$ and $\Psi(\tau, t)$, there exist an infinite number of time functions $\eta(t)$ which will give $\underline{\varphi}(t) = \underline{\xi}$. Thus, one has that

$$E(A\underline{\varphi}(t)) = E(A)E(\underline{\varphi}(t)) \qquad (16)$$

$$E(\langle \underline{x}^*(t), \underline{\varphi}(t) \rangle) = \langle E(\underline{x}^*(t)), E(\underline{\varphi}(t)) \rangle \tag{17}$$

$$E(\langle \underline{p}(t), \underline{\varphi}(t) \rangle) = \langle E(\underline{p}(t)), E(\underline{\varphi}(t)) \rangle \tag{18}$$

Using the relations (13, 16, 17, 18) in Equation (12), and setting the left hand side to zero gives

$$0 = \langle F\bar{\underline{x}}^*(T) - \bar{\underline{p}}(T), \underline{\varphi}(T) \rangle + \int_0^T (\langle Q\bar{\underline{x}}^*(t) + \bar{A}^T\bar{\underline{p}}(t) + \dot{\bar{\underline{p}}}(t),$$

$$\underline{\varphi}(t) \rangle + \eta(t)(\underline{b}^T\bar{\underline{p}}(t) + \lambda u^*(t)))dt \tag{19}$$

In order for Equation (19) to represent an extremum, it is necessary that it be valid with any small perturbation $\underline{\varphi}$ and η. Therefore, it is necessary that:

$$\dot{\bar{\underline{p}}}(t) = -\bar{A}^T\bar{\underline{p}}(t) - Q\bar{\underline{x}}^*(t); \quad \bar{\underline{p}}(T) = F\bar{\underline{x}}^*(T) \tag{20}$$

and

$$u^*(t) = -\frac{1}{\lambda}\underline{b}^T\bar{\underline{p}}(t) \tag{21}$$

Now define $\underline{p}^*(t)$ to be a random process whose mean satisfies Equations (20). When $u^*(t) = -\frac{1}{\lambda}\underline{b}^T\underline{p}^*(t)$, substituting this into Equation (5) and taking expectation will give the following canonnic equations

$$\dot{\bar{\underline{x}}}^*(t) = \bar{A}\bar{\underline{p}}^* - \frac{1}{\lambda}\underline{b}\underline{b}^T\bar{\underline{p}}^*(t); \quad \bar{\underline{x}}^*(0) = \underline{x}_o$$

$$\dot{\bar{\underline{p}}}^*(t) = -Q\bar{\underline{x}}^* - A^T\bar{\underline{p}}^*(t); \quad \bar{\underline{p}}^*(T) = F\bar{\underline{x}}^*(T) \tag{22}$$

Using standard techniques (see, for example, (3)), one can obtain a relationship between $\bar{\underline{p}}^*$ and $\bar{\underline{x}}^*$ of the form

$$\bar{\underline{p}}^*(t) = K(t)\bar{\underline{x}}^*(t) \tag{23}$$

where K(t) satisfies

$$\dot{K}(t) = -K\bar{A} - \bar{A}^TK + \frac{1}{\lambda}K\underline{b}\underline{b}^TK - Q \tag{24}$$

with $K(T) = F$

Then, the optimal control at $t = 0$ is

$$u^*(0) = -\frac{1}{\lambda}\underline{b}^TK(0)\underline{x}_o$$

For the i^{th} realization of the plant, at $t = \epsilon$, there is a new value of the state given by $x(\epsilon) = x_i^*(\epsilon)$. The deviation $\varphi_i(\epsilon) = x_i^*(\epsilon) - \bar{x}^*(\epsilon)$ is due solely to the deviation of $A_i(t)$ from $\bar{A}(t)$ in the interval $[0,\epsilon]$, but the Riccati matrix, $K(t=0)$ was computed for $\bar{A}(t)$ since the exact matrix time function

$A_i(t)$ cannot be known a priori. However, since the plant state is measured, one can compute the matrix $K(t=\epsilon)$ using the same information as before (namely, $\bar{A}(t)$), and realize the control from

$$u^*(t=\epsilon) = -\frac{1}{\lambda}\underline{b}^T K(t=\epsilon)\underline{x}_i^*(\epsilon) \tag{25}$$

while considering the state $\underline{x}_i(t=\epsilon)$ to be the new initial condition.

Thus, (25) is a feedback control law which will give the mean optimal trajectory in the deterministic case when the variance of each of the random parameters in A is zero.

IV. AN EXAMPLE

Description of the Problem Chosen

In this example, several different F and Q matrices will be used in the quadratic cost functional, and the trajectory of the plant for each of the control policies used is shown.

The example is

$$\ddot{y} + a\dot{y} + 2y = u(t) \tag{26}$$

where a is a random variable with the following statistics

$$E(a) = \bar{a} = 3$$

$$E(a(t)a(t+\tau)) = R_a(\tau) = 9 + 3e^{-\frac{|\tau|}{.01}} \tag{27}$$

The state variables chosen are $x_1 = y$ and $x_2 = \dot{y}$ so that (26) can be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) \tag{28}$$

with a cost functional of the form (6) using $\lambda = .1$.

Two cases were considered for this example. For the first case, $Q = 45I$ and $F = 90I$; for the second, $Q = 0$ and $F = 90I$. Figures 1 and 2 show the mean optimal trajectory as well as one computer realization of the plant for each case. As one possible suboptimal control law, the steady-state value of the Riccati matrix K was used as the constant feedback gain. The mean trajectories and the trajectories for a computer realization of the plant are shown in Figures 3 and 4.

The values of the cost functionals for both the mean trajectories and the computer realizations are given under each figure. It is seen that only a small difference exists between the cost with the mean trajectory and for the computer realization of the plant using the optimal control law, but the suboptimal law causes both the mean trajectory and the computer realization of the plant to generate a significantly higher value for the cost functional.

V. CONCLUSIONS

Random parameter systems of the form (1) where it is desired to find the plant output statistics for a given set of parameter statistics and a given plant input have been studied. It has been shown that to obtain an optimal control law for the system which will minimize the expectation of a quadratic cost functional of the form (6), a Riccati equation is derived using the mean of the plant parameters. This Riccati equation will give the required feedback gain to optimize the system.

1. Ariaratnam, S. T., & Loh, N. K., "Optimal Control of Linear Stochastic Systems", Int. J. Control, 6(1), pp. 51-64, 1967.
2. Astrom, K. J., "On a First Order Stochastic Differential Equation", Int. J. Control, 1, pp. 301-326, 1965.
3. Athans, M., & Falb, P., Optimal Control, McGraw-Hill, 1966.
4. Bar-Shalom, I., & Sivan, R., "On the Optimal Control of Discrete-Time Linear Systems with Random Parameters", IEEE G-AC, p. 3, February, 1969.
5. DeRusso, Roy & Close, State Variables for Engineers, John Wiley & Sons, 1967.
6. Drenick, R., & Shaw, L., "Optimal Control of Linear Plants with Random Parameters", IEEE G-AC, pp. 236-244, July, 1964.
7. Dreyfus, S., "Some Types of Optimal Control of Stochastic Systems", SIAM J. on Control Ser. A, 2(1), pp. 120-135, 1964.
8. Forsythe & Wasow, Finite-Difference Methods for Partial Differential Equations, McGraw-Hill.
9. Gunckel, T., & Franklin, G., "A General Solution for Linear Sampled-Data Control", ASME J. Basic Engr. 55, p. 197, 1963.
10. Gururaja, H. V., "Optimal Control of Linear Systems with a Random Parameter", Ph.D. thesis, Univ. of Minnesota, 1969.
11. Joseph, P., "A Study of Digital Adaptive Control Systems", Ph.D. thesis, Purdue University, 1961.
12. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1965.
13. Sworder, D., "Feedback Control of a Class of Linear Systems with Jump Parameters", IEEE G-AC, Feb., 1969.
14. Wonham, M., "Optimal Stationary Control of a Linear System with State-Dependent Noise", SIAM J. on Control, 5(3), pp. 486-500, 1967.
15. Wonham, M., "On a Matrix Riccati Equation of Stochastic Control", SIAM J. on Control, 6(4), pp. 681-697, 1968.
16. Zadicario, I., & Sivan, R., "The Optimal Control of Linear Systems with Unknown Parameters", IEEE G-AC, pp. 423-426, July, 1966.

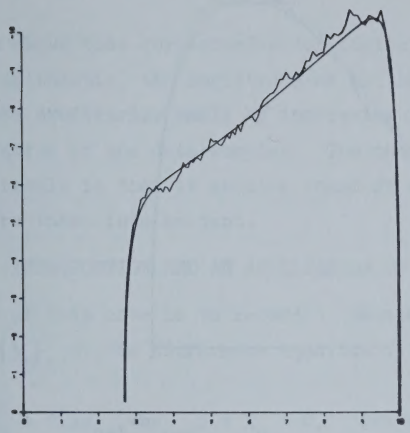


FIGURE NO. 1

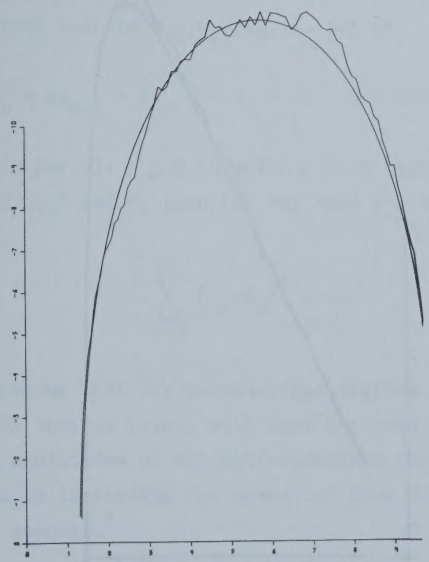


FIGURE NO. 2

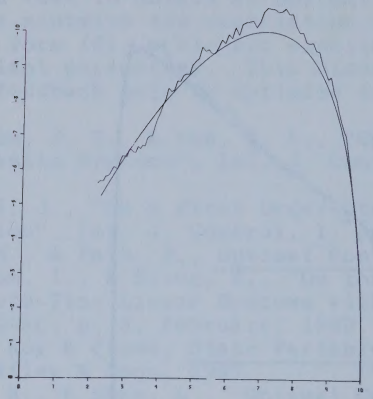


FIGURE NO. 3

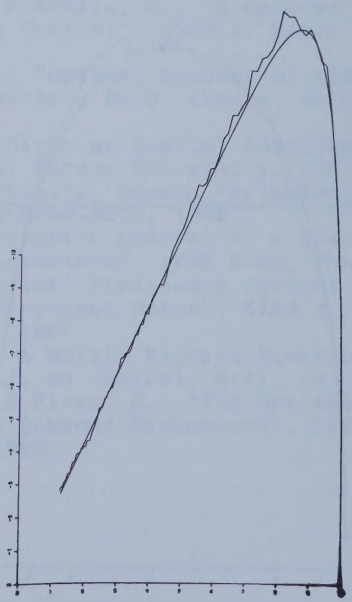


FIGURE NO. 4

A THEOREM CONCERNING LIMIT CYCLES IN DIGITAL FILTERS

I. W. SANDBERG
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

ABSTRACT

The theorem shows that for second-order digital filters employing a certain type of arithmetic, the amplitudes of all self-sustained limit cycles can be made arbitrarily small by increasing the number of bits used in the representation of the data samples. The theorem differs from an earlier related result in that it enables roundoff and a certain additional complication to be taken into account.

I. INTRODUCTION AND AN APPLICATION OF THE THEOREM

The purpose of this note is to record a theorem concerning properties of the solution $\{y_n\}_0^\infty$ of the difference equation

$$y_n + f(ay_{n-1}+by_{n-2}) + \lambda_n = 0, \quad \text{for all } n \geq 0 \tag{1}$$

in which a, b, the y_n, and the λ_n are real numbers, and $f(\alpha) = \alpha$ for all $\alpha \in [-1,1]$ and $f(\alpha) = \text{sign}(\alpha)$ for all real α such that $|\alpha| > 1$. It is proved (see Section II) that if $z^2 + az + b \neq 0$ for all complex z such that $|z| \geq 1$, if the real numbers $w_{-2}, w_{-1}, w_0, \ldots$ satisfy

$$w_n + aw_{n-1} + bw_{n-2} + \lambda_n = 0, \quad \text{for all } n \geq 0, \tag{2}$$

and if $|w_n+\lambda_n| \leq \rho$ for all $n \geq 0$ in which ρ is a certain positive constant that depends only on a and b, then for any real y_{-1} and y_{-2},

$$\sum_{n=0}^{\infty} (y_n-w_n)^2 < \infty. \tag{3}$$

The theorem shows that for second-order digital filters employing a certain type of arithmetic (i.e., with what has been called saturation arithmetic), the amplitudes of all self-sustained limit cycles can be made arbitrarily small by increasing the number of bits used in the representation of the data samples.*

* It is known that this type of proposition cannot be proved when two's-complement arithmetic (i.e., when two's-complement adders) are employed.[1,2]

More explicitly, refering to (1), let $\{y_n\}$ denote a computed sequence, let -a and -b denote machine coefficients, and let the function $f(\cdot)$ and the sequence $\{\lambda_n\}$ model an adder with a finite number of output values $\{v_j\}$ such that $|v_j| \leq 1$ for all j and $|v_j| = 1$ for at least one and at most two values of j. Assume that when the sum of the two inputs to the adder is greater than unity, then its output is $\max_j v_j$, and when the sum of the inputs is less than -1, its output is $\min_j v_j$. The sequence $\{\lambda_n\}$ takes into account the effects of roundoff and the possibility that $\min_j v_j \neq -\max_j v_j$.* By increasing the number of bits used in the representation of the elements of the computed sequence, it is clear that we can make the least upper bound λ of the numbers $|\lambda_0|, |\lambda_1|, \ldots$ as small as we please. Thus for any positive constant ε, we can make λ so small that, with $w_{-2} = w_{-1} = 0$, $|w_n+\lambda_n| \leq \rho$ for all $n \geq 0$ and $|w_n| \leq \frac{1}{2}\varepsilon$ for all $n \geq 0$ [see (2)]. But then, by (3), $(y_n-w_n) \to 0$ as $n \to \infty$. Therefore, given any positive constant ε, we can make λ so small that for some positive integer $\bar{n}$, $|y_n| \leq \varepsilon$ for all $n \geq \bar{n}$.†

II. THE THEOREM

In this section the following result is proved.

Theorem: If, with $f(\cdot)$ as defined in Section I,

(i) a and b are real numbers such that $z^2 + az + b \neq 0$ for all complex z with the property that $|z| \geq 1$,

(ii) $\{\lambda_n\}_0^\infty$ and $\{w_n\}_{-2}^\infty$ are real sequences such that

$$w_n + aw_{n-1} + bw_{n-2} + \lambda_n = 0, \quad n \geq 0$$

with $|w_n+\lambda_n| \leq (1 + |b|)^{-1}(1 - |b|)$ for $n \geq 0$, and

---

* We have $\min_j v_j \neq -\max_j v_j$ whenever two's-complement representation is used for the data, since then for some positive integer t, each data sample is represented by a sum $-\delta_0 + \sum_{k=1}^{t} \delta_k 2^{-k}$ in which each $\delta_k \in \{0,1\}$.

† If we assume that $\min_j v_j = -\max_j v_j$ and that there is no roundoff, then $\lambda_n = 0$ for all $n \geq 0$, and we have $y_n \to 0$ as $n \to \infty$. This was proved earlier.[2]

(iii) $\{y_n\}_{-2}^{\infty}$ is a real sequence such that

$$y_n + f(ay_{n-1}+by_{n-2}) + \lambda_n = 0, \quad n \geq 0$$

then

$$\sum_{n=0}^{\infty} (y_n - w_n)^2 < \infty.$$

Proof:

Let k be an arbitrary nonnegative integer. Consider the linear difference equation

$$u_{n+k} + au_{n+k-1} + bu_{n+k-2} = 0 \quad \text{for all } n \geq 0. \tag{4}$$

With $U_k(z)$ the z-transform of $\{u_{n+k}\}_{n=0}^{\infty}$, we have

$$U_k(z) = u_k H(z) - bu_{k-1} z^{-1} H(z) \tag{5}$$

in which $H(z) = (1+az^{-1}+bz^{-2})^{-1}$. Thus

$$u_{n+k} = u_k h_n - bu_{k-1} h_{n-1}, \quad n \geq 0 \tag{6}$$

where $h_{-1} = 0$ and $\{h_n\}_0^{\infty}$ is the inverse z-transform of $H(z)$.

Using (6),

$$\sum_{m=k}^{\infty} u_m^2 = u_k^2 \sum_{n=0}^{\infty} h_n^2 - 2u_k u_{k-1} b \sum_{n=0}^{\infty} h_n h_{n-1} + b^2 u_{k-1}^2 \sum_{n=0}^{\infty} h_n^2.$$

Let

$$\rho_k[u_k, u_{k-1}] \triangleq \sum_{m=k}^{\infty} u_m^2.$$

Obviously,

$$\rho_{k+1}[u_{k+1}, u_k] = \rho_k[u_k, u_{k-1}] - u_k^2$$

for all $k \geq 0$. The function $\rho_k[\cdot,\cdot]$ possesses the property that

$$\rho_k[q+\delta,u_{k-1}] \geq \rho_k[\text{sign}(q) + \delta,u_{k-1}] \tag{7}$$

whenever $|q| \geq 1$, $|\delta| \leq \varepsilon$, and $|u_{k-1}| \leq 1 + \varepsilon$, in which $\varepsilon = (1 + |b|)^{-1}(1 - |b|)$. This can be shown as follows. Assume that $|q| \geq 1$, $|\delta| \leq \varepsilon$, and $|u_{k-1}| \leq 1 + \varepsilon$. Then

$$\rho_k[q+\delta,u_{k-1}] - \rho_k[\text{sign}(q) + \delta,u_{k-1}]$$

$$= \{(q+\delta)^2 - [\text{sign}(q) + \delta]^2\} \sum_{n=0}^{\infty} h_n^2 - 2[q - \text{sign}(q)]u_{k-1}b \sum_{n=0}^{\infty} h_n h_{n-1}.$$

Since, by the Schwarz inequality,

$$\left| \sum_{n=0}^{\infty} h_n h_{n-1} \right| \leq \sum_{n=0}^{\infty} h_n^2,$$

we have

$$\rho_k[q+\delta,u_{k-1}] - \rho_k[\text{sign}(q) + \delta,u_{k-1}]$$

$$\geq \{(q+\delta)^2 - [\text{sign}(q) + \delta]^2 - 2|q - \text{sign}(q)|(1+\varepsilon)|b|\} \sum_{n=0}^{\infty} h_n^2. \tag{8}$$

The right side of (8) is equal to

$$|q - \text{sign}(q)|[|q| + 1 + 2\delta\ \text{sign}(q) - 2(1+\varepsilon)|b|] \sum_{n=0}^{\infty} h_n^2$$

which is not negative.

With $\{w_n\}_{-2}^{\infty}$ and $\{y_n\}_{-2}^{\infty}$ such that

$$w_n + aw_{n-1} + bw_{n-2} + \lambda_n = 0, \quad n \geq 0$$

and

$$y_n + f(ay_{n-1}+by_{n-2}) + \lambda_n = 0, \quad n \geq 0$$

let $v_n = (y_n - w_n)$ for all $n \geq 0$. Of course

$$v_n + f(ay_{n-1}+by_{n-2}) - aw_{n-1} - bw_{n-2} = 0, \quad n \geq 0. \tag{9}$$

For each $k \geq 0$, the number $\rho_k[v_k, v_{k-1}]$ is well defined. Let us consider the relation between $\rho_k[v_k, v_{k-1}]$ and $\rho_{k+1}[v_{k+1}, v_k]$ for an arbitrary $k \geq 1$. Refering to the discussion concerning (4), we know that

$$\rho_{k+1}[\hat{v}_{k+1}, v_k] = \rho_k[v_k, v_{k-1}] - v_k^2$$

in which $\hat{v}_{k+1} = -(av_k+bv_{k-1})$. Let $q = -(ay_k+by_{k-1})$ and $\delta = (aw_k+bw_{k-1})$. Then $\hat{v}_{k+1} = q + \delta$. From (9), if $|q| \leq 1$, then $v_{k+1} = \hat{v}_{k+1}$, while if $|q| \geq 1$, then $v_{k+1} = \text{sign}(q) + \delta$.

Therefore, by (7), we have

$$\rho_{k+1}[v_{k+1}, v_k] \leq \rho_k[v_k, v_{k-1}] - v_k^2 \tag{10}$$

for all $k \geq 1$. From (10) we have

$$\rho_k[v_k, v_{k-1}] \leq \rho_1[v_1, v_0] - \sum_{n=1}^{(k-1)} v_n^2$$

for all $k > 1$. But $\rho_k[v_k, v_{k-1}] \geq 0$ for all $k > 1$. Thus

$$\sum_{n=1}^{(k-1)} v_n^2 \leq \rho_1[v_1, v_0]$$

for all $k > 1$, which completes the proof of the theorem.

Remark:

It is a simple matter to show that the constant $(1 + |b|)^{-1}(1 - |b|)$, which appears in the statement of the theorem, can be replaced with (the sometimes very much larger number) $(1 + |c|)^{-1}(1 - |c|)$ in which

$$c = b \frac{\sum_{n=0}^{\infty} h_n h_{n-1}}{\sum_{n=0}^{\infty} h_n^2}.$$

REFERENCES

1. Jackson, L. B., "Special-Purpose Hardware for Digital Signal Processing," to be presented at the 1969 NEREM Conference.
2. Ebert, P. M., Mazo, J. E., and Taylor, M. G., "Overflow Oscillations in Digital Filters," to be published in the B.S.T.J. (November 1969).

AN ANALYSIS OF LIMIT CYCLES DUE TO MULTIPLICATION ROUNDING IN RECURSIVE DIGITAL (SUB)FILTERS*

LELAND B. JACKSON
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

ABSTRACT

An analysis of the roundoff limit cycles occuring in fixed-point implementations of second-order recursive digital (sub)filters is presented. A heuristic approach employing an "effective-value" linear model is used to derive certain second-order "deadbands" in which these limit cycles can exist. The results agree with experimental evidence.

INTRODUCTION

Methods for determining whether or not a linear time-invariant sampled-data system is stable are well-known, and they closely parallel the corresponding methods for linear time-invariant continuous systems.[1] However, the implementation of sampled-data systems with digital hardware implies inherent amplitude quantization and a finite dynamic range which introduce nonlinearities into the otherwise linear systems. These digital systems and, in particular, recursive digital filters can thus possess instabilities which are not explained by the usual linear analysis.

Stability in the bounded-input, bounded-output sense is an inherent property of most digital systems because of their finite dynamic range. On the other hand, such systems often do not possess asymptotic stability because periodic outputs, or "limit cycles," can be maintained with zero input to the system. Note that outputs which do not go to zero with zero input must be periodic in a deterministic digital system because of the finite number of internal states for the system.

Two different types of limit cycles can occur in recursive digital filters. The first type is essentially a result of the finite dynamic range as implemented in certain arithmetic systems such as two's-complement binary.[2] These "overflow" limit cycles are generally very large and undesirable, and hence it is important to eliminate the possibility that they can occur. This can always be done, in fact, since one sure way is to zero the internal states of the filter whenever an undesirable overflow occurs. (All overflows are not necessarily undesirable in digital filters as, for example, in the partial sums of more than two data samples).[3] Sandberg has recently shown that a "saturating" form of arithmetic also eliminates overflow limit cycles, permitting only the second type described below.[4]

The second type of limit cycle in recursive digital filters, and the principal subject of this paper, results from amplitude quantization within the filter or, in particular, from the rounding of multiplication products in a feedback loop. Such limit cycles were first noted by R. B. Blackman,[5] who referred to the amplitude intervals within which these limit cycles are

---

* This paper is taken in part from a thesis to be submitted in partial fulfillment of the requirements for the degree of Doctor of Science in the Department of Electrical Engineering at Stevens Institute of Technology.[7]

confined as "deadbands." Blackman considered only first-order (DC) limit cycles, however, (in filters of arbitrary order) with rounding after the summation of the full-accuracy multiplication products. In this paper, we investigate both the first- and second-order deadbands produced in second-order digital (sub)filters as a result of rounding after multiplication. We consider only the zero-input case although results of the same general character can be derived for any constant input. Our investigation is also confined to fixed-point arithmetic implementation because roundoff limit cycles may generally be neglected with floating-point.[6]

COMPLEX-CONJUGATE EFFECTIVE POLES

For a first-order recursive digital (sub)filter with input sequence $\{x_n\}$ and output sequence $\{y_n\}$, the output sequence is assumed to be computed from

$$y_n = x_n - (\alpha y_{n-1})' \tag{1}$$

where $(\ )'$ denotes the operation of rounding to the nearest integer. The first-order deadband within which limit cycles can exist with zero input to the filter is then[5]

$$D_1 = [-k,k] \tag{2a}$$

where k is the largest integer satisfying

$$k \leq \frac{0.5}{1 - |\alpha|} . \tag{2b}$$

The limit cycles occurring within D_1 can be of only two forms: constant magnitude and sign for α negative, or constant magnitude with alternating signs for α positive.

Note that for all output values y_n within D_1, the multiplier α has an "effective value" (α') of ± 1. That is,

$$(\alpha y_n)' = \alpha' y_n = \pm y_n \quad \text{for all } y_n \in D_1. \tag{3}$$

But for a linear discrete filter with transfer function $1/(1+\alpha' z^{-1})$, $\alpha' = \pm 1$ implies that the z-plane pole occurs at $z = \mp 1$. Therefore, we will say that the limit cycles occurring with D_1 are the result of an "effective pole" at $z = \pm 1$.

Now consider the second-order digital (sub)filter defined by

$$y_n = x_n - (\beta_1 y_{n-1})' - (\beta_2 y_{n-2})'. \tag{4}$$

With $\beta_1^2 \leq 4\beta_2$ (complex-conjugate poles) and without rounding, this filter would be stable if, and only if, $0 < \beta_2 < 1$. The case of $\beta_2 = 1$ (i.e., poles on the unit circle in the z plane) is sometimes referred to as being "marginally stable" because with zero input the filter will yield a sinusoidal (and hence nonincreasing) output. But considering a deadband such as that in (2) with $|\alpha|$ replaced by β_2, we see that for all y_n within this deadband, β_2 will have an effective value (β_2') of 1! That is, defining

$$D_2 = [-k,k] \tag{5a}$$

where k is the largest integer satisfying

$$k \leq \frac{0.5}{1-\beta_2}, \quad (0 < \beta_2 < 1), \tag{5b}$$

we then have, as in (3), that

$$(\beta_2 y_n)' = \beta_2' y_n = y_n \quad \text{for all } y_n \in D_2. \tag{6}$$

We thus hypothesize that second-order roundoff limit cycles are a result of an effective complex-conjugate pole pair on the unit circle in the z plane and, as such, are confined (approximately) to the deadband D_2 defined in (5). Although this is an heuristic approach and bounds on the second-order limit cycles have not been derived, extensive computer simulation has corroborated these results in all their essential detail. The only variations from the predicted behavior are the (small) expected deviations from a perfectly sinusoidal waveform due to the rounding errors in the $\beta_1 y_{n-1}$ multiplications. In particular, some of the observed limit cycles for $\beta_2 \geq 0.9$ have exceeded the limits of D_2 by ±1 (i.e., one quantization step), but never more.

Given the amplitude of any particular limit cycle within D_2, the oscillation frequency of the limit cycle can be controlled only by β_1 since the effective value of β_2 is always ±1. To gain insight into this mechanism, it is useful to consider a diagram of the parameter space (β_1,β_2 plane) for the second-order section as shown in Fig. 1. Here, within the (triangular) region of linear stability for the digital filter, the lines $\beta_2 = 0.5$, 0.75, 0.833, and 0.875 have been drawn corresponding to the minimum β_2 values which produce the deadbands [-k,k], k = 1, 2, 3, and 4, respectively. In particular, no limit cycles resulting from complex effective poles are possible for $\beta_2 < 0.5$. The minimum β_2 values (β_{2k}) for the deadband $D_2 = [-k,k]$ are, in general, given by

$$\beta_{2k} = \frac{2k-1}{2k}, \quad k = 1,2,\ldots \tag{7a}$$

Hence, a sufficient condition for a deadband of [-k,k] is simply

$$\beta_2 \geq \beta_{2k}. \tag{7b}$$

Above each horizontal line ($\beta_2 = \beta_{2k}$) in Fig. 1 are drawn the vertical lines $\beta_1 = m(1-\beta_{2k})$ for m odd. Each subregion of the β_1, β_2 plane delimited by these lines must have a common limit cycle behavior because all integers y_n in the corresponding deadband will yield the same rounded products for any β_1, β_2 pair in the subregion. Thus, for example, for $0.5 \leq \beta_2 < 0.75$ corresponding to a deadband of [-1,1], only the subregions $|\beta_1| \geq 1.5$, $0.5 \leq |\beta_1| < 1.5$, and $|\beta_1| < 0.5$ can display distinctive limit cycle behavior because the integers (-1, 0, and +1) contained in this deadband yield the same rounded products for all β_1 values in each of these subregions.

For each of the β_1, β_2 subregions described above, the effective β_1 value (β_1') will be determined by the oscillation frequency of the limit cycle corresponding to that subregion. That is, for an oscillation frequency of f_0 and a Nyquist frequency of f_N (one-half the sampling rate), the effective β_1 value is defined by

$$\beta_1' = -2\cos(\pi f_0/f_N). \tag{8}$$

Thus, to estimate f_0 for a given subregion, we can choose a suitable β_1' and compute the estimate from (8). For example, for the case of $0.5 \leq \beta_2 < 0.75$ discussed above, the effective values β_1' for the subregions $|\beta_1| \geq 1.5$, $0.5 \leq |\beta_1| < 1.5$, and $|\beta_1| < 0.5$ must obviously be just ± 2, ± 1, and 0, respectively. Hence, the oscillation frequencies (f_0) for these subregions are as follows: f_N ($\beta_1' = 2$); $2f_N/3$ ($\beta_1' = 1$); $f_N/2$ ($\beta_1' = 0$); $f_N/3$ ($\beta_1' = -1$); and DC, or $f_0 = 0$ ($\beta_1' = -2$).

The correct β_1' for each subregion is not generally as obvious as it was above, but as β_2 approaches 1, the accuracy of any guess for β_1' must become better and better. Thus, for sufficiently high-Q filters (how high depends upon the application), the approximation $\beta_1' \approx \beta_1$ should suffice to estimate the oscillation frequency f_0. The experimentally determined oscillation frequencies for all subregions having $\beta_2 < .875$ are listed in Tables 1A and 1B. The entries in these tables are for the parameter λ defined by $f_0 = \lambda f_N$. Note that λ must be a rational fraction because, by definition, a discrete limit cycle must be periodic with period NT, where T is the sampling period and N is an integer.

REAL EFFECTIVE POLES

The results in (1)-(3) are easily generalized to include first-order limit cycles in digital filters of arbitrary order with rounding after multiplication. Assuming that

$$y_n = x_n - \sum_{i=1}^{N} (\beta_i y_{n-i})', \tag{9}$$

constant (DC) limit cycles of amplitude k can occur if, and only if,

$$- \sum_{i=1}^{N} a^i (\beta_i k)' = k \tag{10}$$

for $a = +1$; while alternating-sign limit cycles result if, and only if, (10) is satisfied for $a = -1$. In particular, in second-order digital (sub)filters, we must have

$$- (\beta_2 k)' \pm (\beta_1 k)' = k \tag{11a}$$

or, from the definition of the corresponding effective values, simply

$$- \beta_2' \pm \beta_1' = 1. \tag{11b}$$

The lines $- \beta_2 \pm \beta_1 = 1$ form the "sides" of the triangular region of linear stability in Fig. 1, and for any β_1, β_2 pair on these boundaries, at least one of the corresponding z-plane poles must lie at $z = \pm 1$. We will thus say that limit cycles satisfying (11) are due to a real effective pole at $z = \pm 1$.

The subregions of the β_1, β_2 plane which satisfy (11) for a given value of k are readily determined. For $k = 1$, for example, we must have $\pm (\beta_2)' - (\beta_1)' = 1$, which yields the following subregions: $|\beta_1| \geq 0.5$, $|\beta_2| < 0.5$; $|\beta_1| \geq 1.5$, $\beta_2 \geq 0.5$; and $|\beta_1| < 0.5$, $\beta_2 \leq -0.5$. These subregions along with those for $k = 2$ and $k = 4$ are shown in Fig. 2. The

cross-hatched region of Fig. 2 contains those β_1, β_2 points which do not satisfy (11) for any nonzero value of k.

A necessary condition on the β_1, β_2 values satisfying (11) for a given k is seen from Fig. 2 to be

$$|\beta_1| \geq \left(\frac{k-1}{k}\right) + \beta_2. \tag{12}$$

A dotted line in Fig. 2 shows this condition for k = 2. Note that for β_1, β_2 pairs not satisfying (12), the sufficient condition in (7) for a deadband of [-k,k] must also be necessary (remembering that our definition of the deadband D_2 in (5) neglects the possibility of slightly larger limit cycles due to rounding errors from the β_1 multiplication).

UNIFIED RESULTS FOR SECOND-ORDER SECTIONS

We now combine the results of the preceding sections for complex-conjugate and real effective poles, respectively, into a unified description of the limit-cycle behavior of second-order digital filter sections. First, combining Figs. 1 and 2 results in the overall description in Fig. 3. The cross-hatched region in Fig. 3 indicates those β_1, β_2 pairs for which no limit cycles are possible. The subregions corresponding to the deadbands [-1,1] [-2,2], and [-4,4] are also shown. Based upon our heuristic derivation and the supporting experimental results, we hypothesize that a diagram such as Fig. 3 depicts both the necessary and sufficient conditions for a deadband of [-k,k] for all $k \geq 1$ and all β_1 and β_2.

In the case where the filter has complex-conjugate poles (i.e., $\beta_1^2 \leq 4\beta_2$), an alternate depiction of the limit-cycle conditions in terms of the (ideal) pole positions in the z-plane is sometimes more useful. To map the limit-cycle conditions from the β_1, β_2 plane to the z-plane, consider the mapping of the arbitrary line $\beta_2 = s\beta_1 + r$. The complex-conjugate poles of the filter are given by

$$z_p = -\frac{\beta_1}{2} \pm j\sqrt{\beta_2 - \left(\frac{\beta_1}{2}\right)^2} \tag{13}$$

By writing $z_p = x_p + jy_p$, substituting $\beta_2 = s\beta_1 + r$ into (13) and performing some algebra, we arrive at the relationship

$$y_p^2 + (x_p+s)^2 = r + s^2. \tag{14}$$

Thus, the locus of pole positions in the z plane corresponding to the line $\beta_2 = s\beta_1 + r$ for $\beta_1^2 \leq 4\beta_2$ is a circle centered at $z_p = -s$ with a radius of $\sqrt{r+s^2}$.

Applying the above mapping to various lines of interest in Fig. 3 produces the z-plane representation shown in Fig. 4. The necessary conditions for a deadband of [-k,k] are indicated by an N, and the sufficient conditions by an S. These conditions are shown for k = 1, 2, and 4, as before. The separate necessary conditions (circles centered at z = ±1) are taken from (12); while the separate sufficient conditions (vertical lines) are derived as follows: (7) and (11) are sufficient conditions for a deadband of [-k,k]. Therefore, if β_1 satisfies (11) for $\beta_2 = \beta_{2k}$, then either (7) or

(11) must be satisfied for any given value of β_2. The condition on β_1 ensuring that (11) is satisfied for $\beta_2 = \beta_{2k}$ is

$$|\beta_1| \geq \frac{4k-3}{2k} . \tag{15}$$

We now give mathematical expressions for the conditions depicted in Fig. 4 for complex-conjugate poles. First, the sufficient condition (7) becomes

$$|z_p| \geq \sqrt{\frac{2k-1}{2k}} . \tag{16}$$

For $y_p^2 + (|x_p| - 1/3)^2 \geq (2/3)^2$, (16) is also necessary. Otherwise, the necessary condition (12) becomes

$$y_p^2 + (|x_p| - 1)^2 \leq \frac{1}{k} ; \tag{17}$$

while the sufficient condition (15) implies that

$$|x_p| \geq \frac{4k-3}{4k} . \tag{18}$$

Finally, for $|z_p| < 1/\sqrt{2}$ and $k = 1$, the sufficient condition given by (18) is also necessary; that is, for $|z_p| < 1/\sqrt{2}$, a necessary and sufficient condition for a nonzero deadband is $|x_p| \geq \frac{1}{4}$.

CONCLUSIONS

A heuristic approach based upon an "effective value" linear model has been employed to derive the second-order deadbands in which limit cycles due to multiplication rounding in recursive digital filters are found to exist. Extensive computer simulations have been performed for representative β_1, β_2 pairs over the entire region of linear stability for a second-order filter, and these simulations have confirmed the predicted limit-cycle behavior in all cases. The only variations from the predicted results are the (small) expected deviations from a perfectly sinusoidal limit-cycle waveform due to rounding. In particular, some of the observed limit cycles (for $\beta_2 \geq 0.9$) have exceeded the limits of the computed deadband (D_2) by one quantization step, but never more.

Initial experimental evidence also indicated that these results are quite "stable" with respect to the introduction of small nonconstant inputs to the filter. That is, with very small nonconstant inputs, the spectrum of the output remains essentially the same as for the largest zero-input limit cycle. As the input is increased, the output remains largely unchanged (with perhaps some spectral broadening) until the ideal filter output approaches the magnitude of the largest zero-input limit cycle. Above this transition input level, the filter output spectrum approaches the ideal output spectrum more and more closely as the effect of rounding becomes relatively smaller and smaller.

The extension of these results to higher-order digital filters realized as either a parallel or cascade combination of second-order filter sections[3] is obvious in the former case, but not in the latter. In the parallel case, the individual second-order sections share a common input, and their outputs are summed to produce the filter output. Hence, the limit-cycle behavior of each section is independent of all other sections, and the output of the

filter for zero input must be simply the sum of the limit-cycle outputs of the individual sections.

In the cascade case, however, each filter section derives its input from the preceding section, and thus, in the general case, only the first section in the cascade necessarily has zero input with zero input to the filter. The previously described "stability" of our limit-cycle results would suggest that perhaps with small (or zero) inputs the filter behaves essentially as though each section were producing its maximum zero-input limit cycle, or else filtering the output of the previous section, whichever is the larger output. The extent to which this conjecture is valid has yet to be determined.

REFERENCES

1. E. I. Jury, Sampled-Data Control Systems, (New York: John Wiley, 1958).

2. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow Oscillations in Digital Filters," B.S.T.J., Vol. 48, No. 9 (November, 1969).

3. L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An Approach to the Implementation of Digital Filters," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 3 (September, 1968), pp. 413-21.

4. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," Proc. Seventh Annual Allerton Conf. on Circuit and System Theory, Monticello, Illinois (October, 1969).

5. R. B. Blackman, Linear Data-Smoothing and Prediction in Theory and Practice, (Reading, Mass.: Addison-Wesley, 1965), pp. 75-9.

6. I. W. Sandberg, "Floating-Point-Roundoff Accumulation in Digital-Filter Realizations," B.S.T.J., Vol. 46, No. 8 (October, 1967), pp. 1775-91.

7. L. B. Jackson, An Analysis of Roundoff Noise in Digital Filters, Sc.D. Thesis, Stevens Institute of Technology, Hoboken, New Jersey (1969).

TABLE 1A (λ: $f_0 = \lambda f_N$)

$\beta_1 \leq$	$\beta_2 \geq 0.5$ [-1,1]	$\beta_2 \geq 0.75$ [-2,2]	$\beta_2 \geq 0.833$ [-3,3]
0.	1/2	1/2	1/2
-0.167	1/2	1/2	5/11
-0.25	1/2	3/7	4/9
-0.5	1/3	2/5	2/5
-0.75	1/3	1/3	3/8
-0.833	1/3	1/3	1/3
-1.167	1/3	1/3	2/7
-1.25	1/3	1/4	1/4
-1.5	0	1/5	1/5
-1.75	-	0	1/7
-1.833	-	-	0

TABLE 1B (λ: $f_0 = \lambda f_N$)

$\beta_1 \geq$	$\beta_2 \geq 0.5$ [-1,1]	$\beta_2 \geq 0.75$ [-2,2]	$\beta_2 \geq 0.833$ [-3,3]
0.	1/2	1/2	1/2
0.167	1/2	1/2	6/11
0.25	1/2	4/7	5/9
0.5	2/3	3/5	3/5
0.75	2/3	2/3	5/8
0.833	2/3	2/3	2/3
1.167	2/3	2/3	5/7
1.25	2/3	3/4	3/4
1.5	1	4/5	4/5
1.75	-	1	6/7
1.833	-	-	1

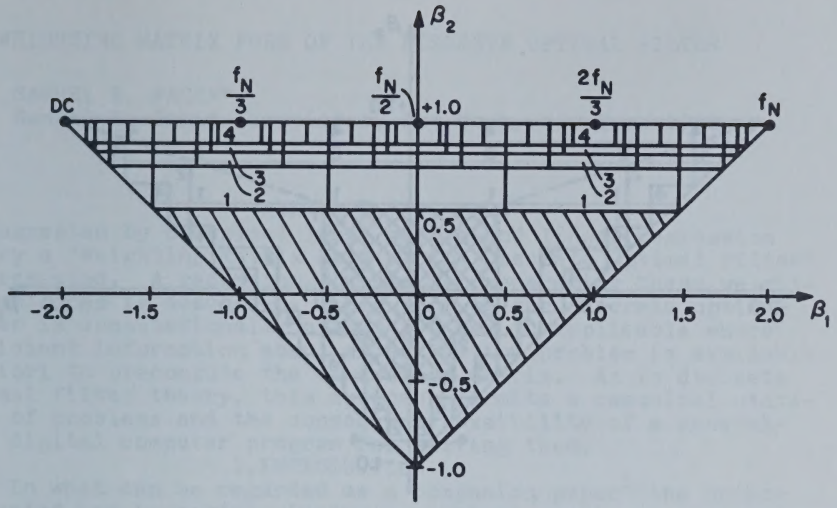


Fig. 1 - Parameter Space for Complex-Conjugate Effective Poles.

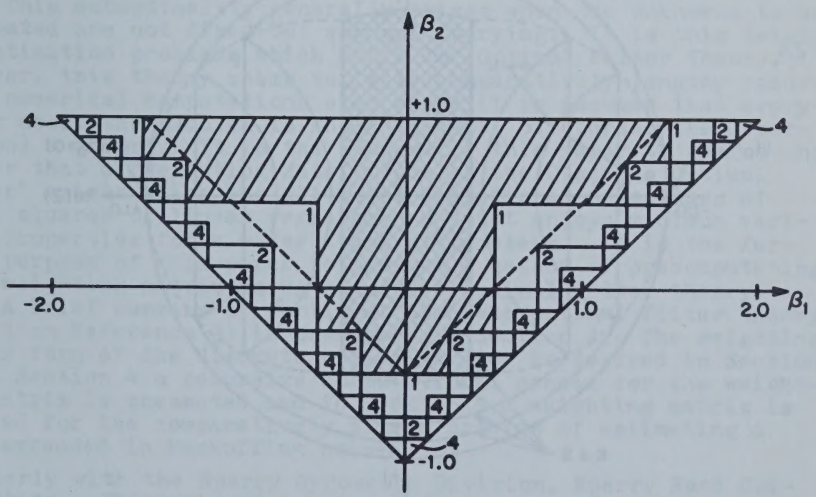


Fig. 2 - Parameter Space for Real Effective Poles.

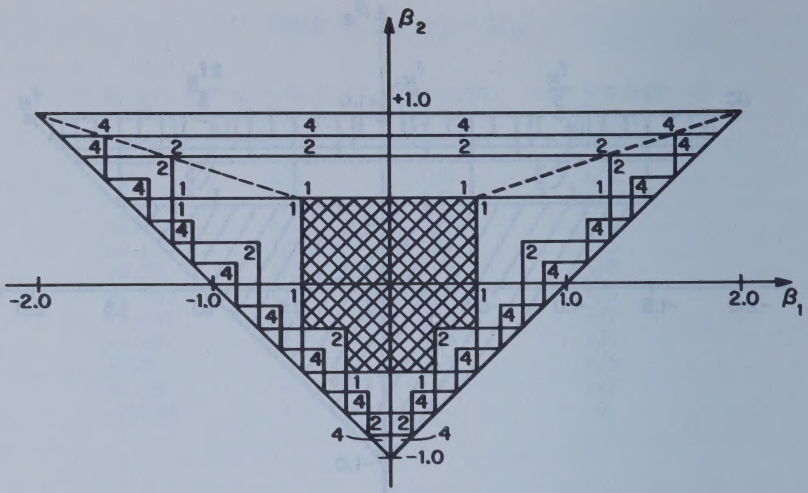


Fig. 3 - All Deadband Subregions (Figs. 1 & 2 combined).

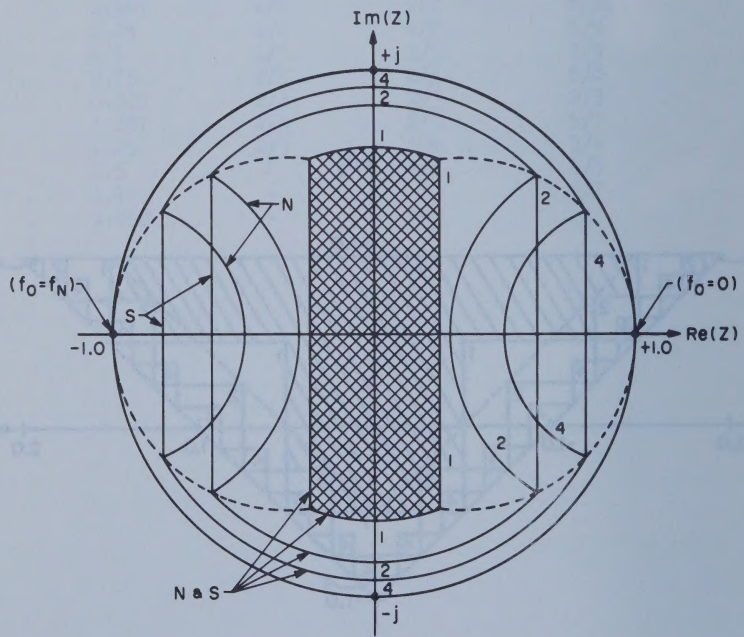


Fig. 4 - Mapping of Fig. 3 to z plane for $\beta_1^2 \leq 4\beta_2$.

THE WEIGHTING MATRIX FORM OF THE DISCRETE OPTIMAL FILTER

SAMUEL L. FAGIN*
Samuel L. Fagin Associates, New York, New York 10024

ABSTRACT

As suggested by classical least square and linear regression theory a "Weighting Matrix Form of the Discrete Optimal Filter" is presented. A recursion formula for generating these weighting matrices is described. This form of the discrete optimal filter is computationally attractive and is applicable where sufficient information about an estimation problem is available a priori to precompute the covariance matrix. As in discrete optimal filter theory, this approach permits a canonical statement of problems and the consequent possibility of a generalized digital computer program for solving them.

1. INTRODUCTION

In what can be regarded as a companion paper[1] the author indicated how in certain instances optimal continuous weighting functions can be generated using optimal filter theory when observation is continuous. This paper is directed towards obtaining analogous "weighting matrices" from discrete optimal filter theory.

The possibility of such a weighting matrix is suggested by the results of classical least squares and/or linear regression. If therein the linear relations between observations and unknowns are known in advance and error statistics are predeterminate the estimates of the unknowns are obtainable as a product of a "weighting matrix" times the vector of observations. The computational simplicity of this approach is so appealing that it is frequently applied even where it is known to be suboptimal in a minimum variance sense.

This suboptimality generally arises when the unknowns to be estimated are not fixed but randomly varying. It is this twist in estimation problems which calls for Optimal Filter Theory.[2,3,4] However, this theory seems to imply comparatively lengthy recursive numerical computations even where it is assumed that everything about the problem is known (except, of course, the observation) a priori. It is the purpose of this paper to remind the reader that a "Weighting Matrix Form of the Discrete Optimal Filter" exists which has all the computational advantages of least squares or linear regression and yet enjoys minimum variance properties for a wider class of problems. It is the further purpose of this paper to present a method of precomputating this weighting matrix using recursive optimal filter theory.

A brief summary of pertinent discrete optimal filter theory (based on Reference 3) is presented in Section 2. The weighting matrix form of the discrete optimal filter is derived in Section 3. In Section 4 a recursive computational scheme for the weighting matrix is presented and in Section 5 a weighting matrix is derived for the comparatively simple problem of estimating a bias embedded in Markoffian noise.

*Formerly with the Sperry Gyroscope Division, Sperry Rand Corporation. The author is grateful for the Sperry support which made this paper possible.

2. SUMMARY OF DISCRETE OPTIMAL FILTER THEORY

For the convenience of the reader, a short summary of discrete optimal filter theory is presented. These results are for a linear continuous system where observations are available at discrete times.

Consider the wide class of continuous systems (or error propagation models thereof) which can be represented by the first-order differential equations

$$\dot{X} = A(t)X + N(t) \tag{2.1}$$

where

X is a k-dimensional column state vector

A(t) is a kXk "system feedback" matrix which may vary with time

N(t) is a k-dimensional column vector of white noise inputs such that their covariance $E[N(t_1)N^T(t_2)] = \pi\delta(t_1-t_2)q(t_1)$ and $E[N(t)] = 0$

δ is the Dirac delta function

$q(t_1)$ is a kXk "noise matrix."

In the modeling adopted here the state vector X includes noise states such as measurement errors and error sources. Most frequently, q is a constant and diagonal, the diagonal elements being the single-ended power spectral density of the white noise necessary to create the noise state of the desired variance.

A sequence of r_i dimensional observations $Y(t_i)$ are made at discrete times t_1, t_2, t_3, . . ., t_i. . . defined by the equation

$$Y(t_i) = M(t_i)X(t_i) \tag{2.2}$$

where $M(t_i)$ is a measurement matrix of dimension r_iXk where $r_i \leq k$. $M(t_i)$ is not necessarily equal to $M(t_j)$ nor does r_i necessarily remain constant.

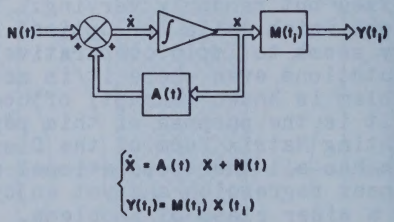


Fig. 1. The canonical model

Equations (1) and (2) can be represented by the multidimensional canonical block diagram of Fig. 1. In this figure and the sequel double lines indicate the flow of vector quantities and matrix boxes multiply their vector inputs from the left.

The optimal filter or estimation problem is to obtain a best estimate for the state vector X based on these observations and any prior knowledge of the states summed up in an initial covariance matrix $P(t_o-0)$.

The solution of this optimal filter or estimation problem is given as follows. The minimum variance estimator for each of the components of X in the class of all unbiased linear estimators is given by the recursive relations.

$$\hat{X}(t_i+s) = \Phi(t_i+s;t_i)\hat{X}(t_i-0) + \Phi(t_i+s;t_i)P(t_i-0)M^T(t_i)$$
$$[M(t_i)P(t_i-0)M^T(t_i)]^{-1}[Y(t_i)-M(t_i)\hat{X}(t_i-0)] \quad (2.3)$$

$$P(t_i+s) = \Phi(t_i+s;t_i)P(t_i-0)\Phi^T(t_i+s;t_i) - \Phi(t_i+s;t_i)P(t_i-0)$$
$$M^T(t_i)[M(t_i)P(t_i-0)M^T(t_i)]^{-1}M(t_i)P(t_i-0)$$
$$\Phi^T(t_i+s;t_i)+Q(t_i+s) \quad (2.4)$$

where

$\hat{X}(t_i+s)$ is the unbiased minimum variance estimator for $X(t_i+s)$

$\hat{X}(t_i-0) = \hat{X}(t_{i-1} + s_{i-1})$ is the unbiased minimum variance estimator for X at time just prior to the measurement, $Y(t_i)$, at time t_i

s_{i-1} is the i-1 interval between observations

$P(t_i+s)$ is the error covariance matrix for the optimal estimate for the state vector at time t_i+s

$P(t_i-0) = P(t_{i-1} + s_{i-1})$ is the covariance matrix at the time just prior to the measurement $Y(t_i)$ at time t_i

$M(t_i)$ is the measurement matrix at time t_i

$\Phi(t_i+s;t_i)$, the state transition matrix, is determined as the solution of the matrix differential equation

$$\frac{d\Phi(t_i+s;t_i)}{ds} = A(t_i+s)\Phi(t_i+s;t_i);\ \Phi(t_i;t_i) = I$$

$$Q(t_i+s) = E[U(t_i+s)U^T(t_i+s)]$$

$$= \pi\Phi(t_i+s;t_i)\int_0^s \Phi^{-1}(t_i+t;t_i)q(t_i+t)\Phi^{-1^T}(t_i+t;t_i)dt$$
$$\Phi^T(t_i+s;t_i)$$

$U(t_i+s)$ is the zero initial condition response to tne white noise input starting at time t_i

$q(t)$ is the noise covariance matrix and is defined by the relation

$E[N(t_1)Nt_2)] = \pi\delta(t_1-t_2)q(t_1)$.

Of course these recursive relations are initiated by an a priori minimum variance estimate of the state vector, $\hat{X}(t_o-0)$, and its corresponding covariance matrix, $P(t_o-0)$, before the first observation is available.

3. THE WEIGHTING MATRIX FORM

A weighting matrix form of discrete optimal filtering is now sought. Define

$$K(t_i) \triangleq P(t_i-0)M^T(t_i)\,[M(t_i)P(t_i-0)M^T(t_i)]^{-1} \quad (3.1)$$

Then Equation(2.3) becomes

$$\hat{X}(t_i+s) = \Phi(t_i+s;t_i)\{[I-K(t_i)M(t_i)]\hat{X}(t_i-0) + K(t_i)Y(t_i)\} \quad (3.2)$$

Assuming an apriori optimal estimate, $\hat{X}(t_o-0)$, and its covariance matrix, $P(t_o-0)$, is available then at time t_o+s_o after the

observation at time t_o

$$\hat{X}(t_o+s_o) \triangleq \hat{X}(t_1-0) = \Phi(t_o+s_o;t_o)$$

$$\{[I-K(t_o)M(t_o)]\hat{X}(t_o-0)+K(t_o)Y(t_o)\} \qquad (3.3)$$

Then after another observation at time t_1

$$\hat{X}(t_1+s_1) = \Phi(t_1+s_1;t_1)\,[I-K(t_1)M(t_1)]$$

$$\Phi(t_o+s_o;t_o)\,[I-K(t_o)M(t_o)]\hat{X}(t_o-0)+\Phi(t_1+s_1;t_1)$$

$$[I-K(t_1)M(t_1)]\Phi(t_o+s_o;t_o)Y(t_o)+\Phi(t_1+s_1;t_1)K(t_1)\,Y(t_1)$$

Again after an observation at time t_2 (3.4)

$$\hat{X}(t_2+s_2) = \Phi(t_2+s_2;t_2)\,[I-K(t_2)M(t_2)]\Phi(t_1+s_1;t_1)$$

$$[I-K(t_1)M(t_1)]\Phi(t_o+s_o;t_o)[I-K(t_o)M(t_o)]\hat{X}(t_o-0)$$

$$+\Phi(t_2+s_2;t_2)[I-K(t_2)M(t_2)]\Phi(t_1+s_1;t_1)$$

$$[I-K(t_1)M(t_1)]\Phi(t_o+s_o;t_o)K(t_o)Y(t_o)+\Phi(t_2+s_2;t_2)$$

$$[I-K(t_2)M(t_2)]\Phi(t_1+s_1;t_1)K(t_1)Y(t_1)$$

$$+\Phi(t_2+s_2;t_2)K(t_2)Y(t_2) \qquad (3.5)$$

and so on. It is clear that

$$\hat{X}(t_n+s_n) = \Phi(t_n+s_n;t_n)[V_n\hat{X}(t_o-0)+\sum_{i=o}^{n} J_{in}K(t_i)Y(t_i)] \qquad (3.6)$$

where

$$J_{in} = [I-K(t_n)M(t_n)]\Phi(t_{n-1}+s_{n-1};t_{n-1})\cdots\cdots$$

$$\cdots\cdots[I-K(t_{i+2})M(t_{i+2})]\Phi(t_{i+1}+s_{i+1};t_{i+1})$$

$$[I-K(t_{i+1})M(t_{i+1})]\Phi(t_i+s_i;t_i) \quad ; \; i \leq n-1 \qquad (3.7)$$

$$J_{nn} = I$$

$$V_n = J_{on}[I-K(t_o)M(t_o)] \qquad (3.8)$$

Define

$$Y_n \triangleq \begin{bmatrix} \hat{X}(t_o-o) \\ \hline Y(t_o) \\ \hline Y(t_1) \\ \hline \vdots \\ \hline Y(t_n) \end{bmatrix} \qquad (3.9)$$

$$W(t_n+s_n) \triangleq \Phi(t_n+s_n;t_n)[V_n \vdots J_{on}K(t_o) \vdots J_{1n}K(t_1) \vdots J_{2n}K(t_2) \vdots \cdots$$

$$\vdots J_{nn}K(t_n)] \qquad (3.10)$$

Then Equation (3.6) can be written

$$\hat{X}(t_n+s_n) = W(t_n+s_n)Y_n \tag{3.11}$$

where Y_n is a vector of dimension $k+\sum_{i=o}^{n} r_i$ and the "Weighting Matrix", $W(t_n+s_n)$ has k rows and $k+\sum_{i=o}^{n} r_i$ columns. This is the "Weighting Matrix Form of the Discrete Optimal Filter".

Where the number of observations is finite and all the information is available to precompute the covariance matrix and hence the "weighting matrix" Equation (3.11) is a computationally very attractive form of the optimal filter estimate for $X(t_n+s_n)$. This is particularly true when only one component of X is to be estimated. In such a case only a single row of $W(t_n+s_n)$ need be stored and the estimate is merely the product of this row times the column vector Y_n.

4. RECURSIVE COMPUTATION OF THE WEIGHTING MATRIX

While it is possible to compute the weighting matrix from the equations of Section 3 these relations do not make it apparent that this weighting matrix can be computed recursively as demonstrated below. Assume that the weighting matrix, $W(t_{n-1}+s_{n-1})$, is available. Thus,

$$\hat{X}(t_{n-1}+s_{n-1}) = W(t_{n-1}+s_{n-1})Y_{n-1} \tag{4.1}$$

and by Equation (3.2)

$$\begin{aligned}\hat{X}(t_n+s_n) &= W(t_n+s_n)Y_n \\ &= \Phi(t_n+s_n;t_n)\{[I-K(t_n)M(t_n)] \\ &\quad W(t_{n-1}+s_{n-1})Y_{n-1}+K(t_n)Y(t_n)\}\end{aligned} \tag{4.2}$$

It is clear from this relation that

$$\begin{aligned}W(t_n+s_n) &= \Phi(t_n+s_n;t_n) \\ &\quad \{[I-K(t_n)M(t_n)]W(t_{n-1}+s_{n-1})\vdots K(t_n)\}\end{aligned} \tag{4.3}$$

This is the recursive relation for the weighting matrix. It can be initiated by

$$W(t_o+s_o) = \Phi(t_o+s_o;t_o)\{[I-K(t_o)M(t_o)]\vdots K(t_o)\} \tag{4.4}$$

Note that Equation (4.3) indicates the tacking on of $K(t_n)$ at each update of the weighting matrix adding "r_i" (the dimension of the ith observation) more columns.

Further note that the substitution of Equation (3.1) into Equation (2.4) results in

$$\begin{aligned}P(t_i+s) &= \Phi(t_i+s;t_i)[I-K(t_i)M(t_i)] \\ &\quad P(t_i-0)\Phi^T(t_i+s;t_i)+Q(t_i+s)\end{aligned} \tag{4.5}$$

This points up the fact that any computer program for generating the covariance matrix of an optimal filter (such as mentioned in Reference 3) would have available the matrices necessary to update the weighting matrix as per Equation (4.3). The generation of the covariance matrix and the weighting matrix could proceed in parallel.

5. SIMPLE EXAMPLE, THE OPTIMAL ESTIMATION OF A BIAS EMBEDDED IN MARKOFFIAN NOISE.

As an illustration of the weighting matrix form of the discrete optimal filter consider the discrete version of the simple problem taken up in Reference 1. This is the problem of obtaining the minimum variance estimate of an unknown bias from a finite number of observations which are contaminated by Markoffian noise. In addition to being familiar through study in Reference 1 a linear regression solution to this particular discrete problem is available. This solution is presented in Section (5.1) for comparison and checking purposes. The generation of the weighting matrix from discrete optimal filter theory is demonstrated in Section (5.2).

5.1 Linear Regression Solution*

A signal, $Y(t)$, consisting of an unknown bias plus Markoffian noise is observed at equally spaced discrete times $t_o, t_1, t_2, \ldots, t_n$. That is

$$Y(t) = C + u(t)$$

and $Y(t_o), Y(t_1), \ldots, Y(t_n)$ is available where

$u(t)$ is Markoffian noise with zero mean and auto-correlation function $\sigma^2 e^{-\frac{|t|}{\tau}}$

C = unknown bias

It is desired to obtain a minimum variance estimator for the unknown bias, C.

It is shown in Appendix A that via Linear Regression the minimum variance estimator in the class of all unbiased linear estimators is given by

$$\hat{C}_n = \frac{Y(t_o)+Y(t_n)+(\frac{1}{\gamma} -1)\sum_{i=o}^{n} Y(t_i)}{2 + (n+1)(\frac{1}{\gamma} - 1)} \tag{5.1}$$

where

$$\gamma \triangleq e^{\frac{-(t_n-t_o)}{n\tau}} \tag{5.2}$$

Further the variance of this estimate is given by

$$P_n = \frac{\sigma^2 (1 + \gamma)}{(n+1)(1-\gamma) +2\gamma} \tag{5.3}$$

and if W_n^* is defined by

$$\hat{C}_n \triangleq [W_n^*] \begin{bmatrix} Y(t_o) \\ Y(t_1) \\ \cdot \\ \cdot \\ \cdot \\ Y(t_n) \end{bmatrix}$$

then

$$W_n^* = \frac{1}{(n+1)(1-\gamma)+2\gamma} [1 \ (1-\gamma)(1-\gamma)..(1-\gamma) \ 1] \tag{5.4}$$

---

* H. Strell obtained this solution in 1962 leaning heavily on the results of Reference 6.

As far as is known by the author this solution is readily available only for equal intervals of time.

5.2. The Optimal Filter Approach

The identical results are now sought via the weighting matrix form of the discrete optimal filter. As in Reference 1 define

$$X = \begin{bmatrix} C \\ u \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{1}{\tau} \end{bmatrix}, \quad N = \begin{bmatrix} 0 \\ n_u \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$q = \begin{bmatrix} 0 & 0 \\ 0 & \frac{2}{\pi}\frac{\sigma^2}{\tau} \end{bmatrix}, \quad \hat{X}(t_o-0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \text{ and } P(t_o-0) = \begin{bmatrix} \Sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

where it is understood that, reflecting complete lack of prior knowledge concerning C, Σ^2 will later be allowed to become infinite. With these definitions the problem of estimating an unknown bias embedded in Markoffian noise can be stated in canonical form (c.f. Equations (2.1) and (2.2)).

$$\dot{X} = AX + N$$

$$Y(t_i) = M(t_i)X(t_i); \quad t_i = t_o, t_1, t_2, \ldots, t_n$$

and the results of Sections 2, 3 and 4 apply. It is now desired to duplicate the results of Section 5.1 using the weighting matrix form of the optimal filter.

First note that by Equations (3.1), (2.4), and (4.5) that

$$K(t_o) = \begin{bmatrix} \frac{\Sigma^2}{\sigma^2+\Sigma^2} \\ \frac{\sigma^2}{\sigma^2+\Sigma^2} \end{bmatrix} \tag{5.5}$$

$$P(t_o+0) = \begin{bmatrix} \frac{\sigma^2\Sigma^2}{\sigma^2+\Sigma^2} & \frac{-\sigma^2\Sigma^2}{\sigma^2+\Sigma^2} \\ \frac{-\sigma^2\Sigma^2}{\sigma^2+\Sigma^2} & \frac{\sigma^2\Sigma^2}{\sigma^2+\Sigma^2} \end{bmatrix} \tag{5.6}$$

As anticipated earlier, let Σ^2 go infinite reflecting complete lack of a priori knowledge of C. Thus

$$K(t_o) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5.7}$$

and

$$P(t_o+0) = \begin{bmatrix} \sigma^2 & -\sigma^2 \\ -\sigma^2 & \sigma^2 \end{bmatrix} \tag{5.8}$$

The A matrix being constant and the interval between observations being fixed the state transition matrix is for all intervals.

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix} \tag{5.9}$$

where $\gamma = e^{\frac{-(t_n-t_o)}{n\tau}}$

Thus by Equation 4.4

$$W(t_o+s_o) = \begin{bmatrix} 0 & -1 & 1 \\ 0 & \gamma & 0 \end{bmatrix} \tag{5.10}$$

where implicitly

$$Y_o = \begin{bmatrix} 0 \\ 0 \\ Y(t_o) \end{bmatrix} \tag{5.11}$$

For this problem

$$Q(t_o+s_o) = Q(t_1+s_1) = \cdots = \begin{bmatrix} 0 & 0 \\ 0 & (1-\gamma^2)\,\sigma^2 \end{bmatrix} \tag{5.12}$$

and using Equation (4.5)

$$P(t_o+s_o) = \begin{bmatrix} \sigma^2 & -\sigma^2\gamma \\ -\sigma^2\gamma & \sigma^2 \end{bmatrix} \triangleq P(t_1-0) \tag{5.13}$$

Substituting this in Equation (3.1)

$$K(t_1) = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \tag{5.14}$$

and

$$[I-K(t_1)M(t_1)] = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{5.15}$$

Thus employing the recursion formula of Equation (4.3) there results

$$W(t_1+s_1) = \begin{bmatrix} 0 & -\frac{(1+\gamma)}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{+(1+\gamma)\gamma}{2} & -\frac{\gamma}{2} & \frac{\gamma}{2} \end{bmatrix} \tag{5.16}$$

Equation (4.5) is now used again to obtain

$$P(t_1+s_1) = \frac{\sigma^2(1+\gamma)}{2} \begin{bmatrix} 1 & -\gamma \\ -\gamma & (\gamma^2-2\gamma+2) \end{bmatrix} = P(t_2-0) \tag{5.17}$$

Repeating the process of the above paragraph

$$K(t_2) = \frac{1}{3-\gamma} \begin{bmatrix} 1 \\ 2-\gamma \end{bmatrix} \tag{5.18}$$

$$[I-K(t_2)M(t_2)] = \frac{1}{3-\gamma} \begin{bmatrix} (2-\gamma) & -1 \\ -(2-\gamma) & 1 \end{bmatrix} \tag{5.19}$$

$$W(t_2+s_2) = \frac{1}{3-\gamma} \begin{bmatrix} 0 & -(1+\gamma) & 1 & 1-\gamma & 1 \\ 0 & (1+\gamma)\gamma & -\gamma & -(1-\gamma)\gamma & (2-\gamma)\gamma \end{bmatrix} \tag{5.20}$$

$$P(t_2+s_2) = \frac{\sigma^2(1+\gamma)}{3-\gamma} \begin{bmatrix} 1 & -\gamma \\ -\gamma & 3-4\gamma+2\gamma^2 \end{bmatrix} \tag{5.21}$$

This process can be repeated over and over. However, for the purposes of illustration it will suffice to observe that the last three terms of the upper row of weighting matrix Equation (5.20) are identical to the weighting matrix arrived at independantly in Section 5 Equation (5.4) where n=2. As $\hat{X}(t_o-0) = 0$ it is seen that the indicated minimum variance estimates for C are identical. Comparison of Equations (5.3) and

(5.21) show that the calculated variances are identical too.

It should be apparent from the above that the Linear Regression approach is much to be preferred over the optimal filter approach for this particular problem. However it doesn't take much of a variation in this problem to make the latter the preferable approach. In any case, it is anticipated that a generalized digital computer program similar to that mentioned in Reference 3 will be available to do the "horse" work.

6. CONCLUSIONS

A weighting matrix form of the discrete optimal filter has been presented. A recursion formula for the weighting matrix has been derived which appears to be relatively simple to compute. It has been shown for a simple specific problem that this weighting matrix form of the optimal filter leads to a solution identical to that arrived at independantly.

These weighting matrices are only applicaple when everything about the estimation problem except the observations are known in advance. While in some special cases (the simple problem of Section 5 is one) optimal weighting matrices are obtainable via Linear Regression Theory, the weighting matrices described herein have a much wider range of applicability with a comparable computational simplicity.

REFERENCES

1. Fagin, S.L., "The Weighting Function Form of the Continuous Optimal Filter," 1968 Joint Automatic Control Conference Record, pp 619-625.
2. Kalman, R.E., "A New Approach to Linear Filtering and Prediction Problems", Journal of Basic Engineering", Vol.82,No.1, March,1960.
3. Fagin,S.L., "Recursive Linear Regression Theory,Optimal Filter Theory, and Error Analyses of Optimal Systems,; 1964 IEEE International Conv. Rec., Vol.12, pt.1.
4. Swerling, P., "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations", The Journal of the Astronautical Sciences, Vol. VI, No. 3, Autumn, 1959.
5. Scheffe, H., The Analysis of Variance, Chapter 1, New York, John Wiley and Sons, 1959.
6. Reich,E. and Swerling, P., "The Detection of a Sine Wave in Gaussian Noise", Journal of Applied Physics,Vol. 24, No.3, 289-296,March 1953.

APPENDIX A

Linear Regression Solution to the Simple Example; The Optimal Estimation of a Bias Embedded in Markoffian Noise.

A signal, Y(t), consisting of an unknown bias plus (exponentially correlated "wide-sense") Markoffian noise is observed at equally spaced discrete times t_o, t_1, t_2, . .,t_n. That is

$$Y(t) = C + u(t)$$

and $Y(t_o)$, $Y(t_1)$, ..., $Y(t_n)$ is available where u(t) is Markoffian noise with zero mean and autocorrelation function

$$\sigma^2 e^{-\frac{|t|}{\tau}}$$

C = unknown bias

It is desired to obtain a minimum variance estimator for the unknown bias C.

Define

$$Y_n \triangleq \begin{bmatrix} Y(t_o) \\ Y(t_1) \\ Y(t_2) \\ \cdot \\ \cdot \\ \cdot \\ Y(t_n) \end{bmatrix}, \quad B_n \triangleq \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \quad \epsilon_n \triangleq \begin{bmatrix} u(t_o) \\ u(t_1) \\ u(t_2) \\ \cdot \\ \cdot \\ \cdot \\ u(t_n) \end{bmatrix}$$

and $X \triangleq C$. Then the problem can be stated as follows: Given

$$B_n X = Y_n - \epsilon_n$$

$$E(\epsilon_n) = 0$$

$$\Sigma_n = E(\epsilon_n \epsilon_n^T) = \sigma^2 \begin{bmatrix} 1 & \gamma & \gamma^2 & \cdot & \cdot & \cdot & \cdot & \gamma^n \\ \gamma & 1 & \gamma & \cdot & \cdot & \cdot & \cdot & \gamma^{n-1} \\ \gamma^2 & \gamma & 1 & \cdot & \cdot & \cdot & \cdot & \gamma^{n-2} \\ \cdot & \cdot & \cdot & \cdot & & & & \cdot \\ & & & & \cdot & & & \\ \cdot & \cdot & \cdot & & & \cdot & \cdot & \\ & & & & & & \cdot & \\ \gamma & \gamma^{n-1} & \gamma^{n-2} & & & & & 1 \end{bmatrix}$$

where $\gamma = e^{-\frac{(t_n - t_o)}{n\tau}}$. Find the minimum variance estimate for X defined equal to C. In accordance with classical linear regression theory[5] the minimum variance estimator in the class of all unbiased linear estimators for X=C is given by

$$\hat{C}_n \triangleq \hat{X}_n = (B_n^T \Sigma_n^{-1} B_n)^{-1} B_n^T \Sigma_n^{-1} Y_n \qquad \text{(A.1)}$$

where the variance of the estimate is given by

$$P_n = (B_n^T \Sigma_n^{-1} B_n)^{-1} \qquad \text{(A.2)}$$

The only remaining obstacle is obtaining Σ_n^{-1}, the inverse of the covariance matrix of errors. For the simple example under consideration this has been given in Reference 6 as

$$\Sigma_n^{-1} = \frac{1}{\sigma^2(1-\gamma^2)} \begin{bmatrix} 1 & -\gamma & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ -\gamma & (1+\gamma^2) & -\gamma & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & -\gamma & (1+\gamma^2) & -\gamma & & & & & & \cdot \\ \cdot & 0 & -\gamma & \cdot & & & & & & \cdot \\ \cdot & & & & \cdot & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & & (1+\gamma^2) & -\gamma & \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & & -\gamma & 1 & \end{bmatrix}$$

For the simple example at hand

$$B_n^T \Sigma_n^{-1} B_n = \text{sum of terms in } \Sigma_n^{-1}$$

$$= \frac{2-2n\gamma+(n-1)(1+\gamma^2)}{\sigma^2(1-\gamma^2)}$$

$$= \frac{(n+1)(1-\gamma)+2\gamma}{\sigma^2(1+\gamma)}$$

Thus

$$P_n = \frac{\sigma^2(1+\gamma)}{(n+1)(1-\gamma)+2\gamma} \qquad \text{(A.3)}$$

the variance of the estimate. Now

$$B_n^T \Sigma_n^{-1} = \frac{1}{\sigma^2(1+\gamma)} \left[1 \ (1-\gamma)(1-\gamma). \ . \ .(1-\gamma) \ 1\right] .$$

Thus from equation (A.1)

$$\hat{C}_n = P_n B_n^T \Sigma_n^{-1} Y_n \qquad (A.4)$$

$$= \frac{1}{(n+1)(1-\gamma)+2\gamma} \left[1 \ (1-\gamma)(1-\gamma). \ . \ .(1-\gamma) \ 1\right] \begin{bmatrix} Y(t_0) \\ Y(t_1) \\ Y(t_2) \\ \cdot \\ \cdot \\ \cdot \\ Y(t_n) \end{bmatrix}$$

and if W_n^* is defined by

$$\hat{C}_n = W_n^* Y_n \qquad (A.5)$$

then

$$W_n^* = \frac{1}{(n+1)(1-\gamma) + 2\gamma} \left[1 \ (1-\gamma)(1-\gamma). \ . \ .(1-\gamma) \ 1\right] \qquad (A.6)$$

Equations A.3, A.4, A.5 and A.6 are the solution for the simple example of estimating a bias embedded in Markoffian noise.

RESPONSE IDENTIFICATION OF DISTRIBUTED SYSTEMS WITH NOISY MEASUREMENTS AT FINITE POINTS*

PAUL C. BADAVAS
The Foxboro Company
Foxboro, Massachusetts

GEORGE N. SARIDIS
Purdue University
Lafayette, Indiana

INTRODUCTION

The method of stochastic approximation [2,3,8] that minimizes a mean-square error criterion is used for the on-line estimation of a set of parameters in an assumed approximate response of a class of distributed systems. The systems of concern in this paper are those that can be modeled mathematically by the general linear elliptic equation on a closed and bounded domain Ω. Moreover, it is assumed that the model is completely or partially unknown. No attempt will be made herein to recover the model. Instead the response of the system is approximated by a finite set of linearly independent and bounded functions and an unknown set of parameters. The type as well as the number of functions used in the approximate solution is assumed determined a priori.

It is assumed that noisy measurements of the response are available at a finite number of fixed points in the domain Ω. The minimum number of measurement points is determined by the number of functions used in the approximation of the response. A method for determination of the minimum number of measurement points necessary for the recovery of the solutions using a n-dimensional extension of the sampling theorem has been developed in [1]. This method is applicable to the present problem.

In the second part of the paper a random optimization algorithm [4,5] is presented in order to determine the location of the finitely many measurement points in Ω such that a sensitivity criterion is maximized. This is done in order to place the measurement devices in the most "sensitive" locations and thus accelerate the convergence of the search algorithm. A positive definite matrix is obtained from the performance criterion in terms of the preselected approximating functions and the measurement points. The minimum eigenvalue of this matrix, which is generally a multimodal surface in the measurement points, is chosen as the sensitivity criterion.

STATEMENT OF THE PROBLEM

Consider a linear distributed system defined on a fixed spatial domain Ω, which is an open connected subset of the N-dimensional Euclidean space E^N. Let Ω_b denote the boundary of Ω. The following elliptic linear partial differential equation describes the system:

$$L\,u\,(X) = f(X) \qquad X \in \Omega \tag{1}$$

with boundary condition:

$$L_1\,u\,(X^b) = g(X^b) \qquad X^b \in \Omega_b \tag{2}$$

L and L_1 are scalar linear differential operators whose parameters may depend upon X and X^b, respectively. More importantly it is assumed that these operators are completely or partially unknown.

*This work was supported by the Joint Services Electronics Program under ARPA Contract No. N00014-67-A0226-0004.

$X^T = [x_1,..,x_N]$ the spatial coordinate vector

$[X^b]^T = [x_1^b,..,x_N^b]$ the boundary spatial coordinate vector

$u(X)$ is a scalar-valued function which is the response of the system defined and bounded for all $X \in \Omega$.

It is assumed that a unique solution of (1) with boundary condition (2) exists.

Let the domain Ω be the following normalized domain:

$$\Omega \triangleq [0 < x_i < \pi \; ; \; i=1,\ldots,N] \tag{3}$$

The solution $u(x_1,..,x_N)$ of equation (1) with boundary condition (2) will be approximated in each of the coordinates x_i; i=1,..,N by a finite set of preselected linearly independent and bounded functions, and an unknown set of parameters multiplying these functions. For example, Fourier series expansions can be used. Let $\psi^i_{k_i}$; $k_i=1,..,K_i$ be the finite set of bounded functions selected for the i^{th} coordinate. Then the approximation of $u(x_1,..,x_N)$ in the domain Ω equation (3) is defined as follows:

$$\hat{u}(x_1,..,x_N) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_N=1}^{K_N} a_{k_1,..,k_N} \psi^1_{k_1}(x_1)..\psi^N_{k_N}(x_N) = c^T \psi(x_1,..,x_N) \tag{4}$$

The vectors c and $\psi(x_1,..,x_N)$ are defined by:

$$c^T \triangleq \left[a_{1,..,1},..,a_{1,..,K_N},..,a_{K_1,K_2,..,1},..,a_{K_1,K_2,..,K_N}\right] \tag{5}$$

$$\psi^T(x_1,..,x_N) \triangleq \left[\psi^1_1(x_1)..\psi^N_1(x_N),..,\psi^1_1(x_1)..\psi^N_{K_N}(x_N),\; \psi^1_1(x_1)..\psi^{N-1}_2(x_{n-1})\right.$$

$$\left.\psi^N_1(x_N),..,\psi^1_{K_1}(x_1)\psi^2_{K_2}(x_2)..\psi^N_1(x_N),..,\psi^1_{K_1}(x_1)\psi^2_{K_2}(x_2)..\psi^N_{K_N}(x_N)\right] \tag{6}$$

The index K is defined as the product:

$$K \triangleq K_1 \cdots K_N \tag{7}$$

Noisy measurements of the response are available at a finite number of points in the domain Ω. That is,

$$v(x_1^{i_1},..,x_N^{i_N}) = u(x_1^{i_1},..,x_N^{i_N}) + \eta(x_1^{i_1},..,x_N^{i_N}); \; i_1=1,..,I_1;..; \; i_N=1,..,I_N \tag{8}$$

It is assumed that the measurement noise η with an unknown continuous probability density $p(\eta)$, satisfies the following conditions:

$$E\{\eta(x_1^{i_1},..,x_N^{i_N})\} = 0 \tag{9}$$

$$E\{\eta^2(x_1^{i_1},..,x_N^{i_N})\} = \sigma^2(i_1,..,i_N) < \infty; \; i_1=1,..,I_1; \; i_N=1,..,I_N$$

$$E\left\{\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\eta^2(x_1^{i_1},..,x_N^{i_N})\right\} = \sigma^2 < \infty \tag{10}$$

where $\sigma^2(i_1,..,i_n)$ and σ^2 are unknown.

The identification of the response $u(x_1,..,x_N)$ consists of finding the parameter vector $c=c^*$ that minimizes the following mean square error criterion:

$$J(c) = \frac{1}{2} E\left\{ \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} [v(x_1^{i_1},..,x_N^{i_N}) - c^T \psi(x_1^{i_1},..,x_N^{i_N})]^2 \right\} \quad (11)$$

Minimizing the above criterion with respect to c yields:

$$\frac{dJ(c)}{\partial c} = - E\left\{ \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} [v(x_1^{i_1},..,x_N^{i_N}) - c^T \psi(x_1^{i_1},..,x_N^{i_N})] \psi(x_1^{i_1},..,x_N^{i_N}) \right\} = 0 \quad (12)$$

Using the fact that the measurement points are fixed, the above relation gives:

$$c^* = \left[\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \psi(x_1^{i_1},..,x_N^{i_N}) \psi^T(x_1^{i_1},..,x_N^{i_N}) \right]^{-1} E\left\{ \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} v(x_1^{i_1},..,x_N^{i_N}) \psi(x_1^{i_1},..,x_N^{i_N}) \right\} \quad (13)$$

provided that the matrix $S(X_i)$ defined below is positive definite. That is:

$$S(X) \triangleq \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \psi(x_1^{i_1},..,x_N^{i_N}) \psi^T(x_1^{i_1},..,x_N^{i_N}) > 0 \quad (14)$$

A necessary condition for the above relation (14) to hold is that the number of measurement points in each coordinate must be equal to or greater than the number of functions used to approximate the response in the same coordinate. That is:

$$I_1 \geq K_1, \cdots, I_N \geq K_N \quad (15)$$

The total number of measurement points is the product:

$$I \triangleq I_1 \cdots I_N \quad (16)$$

Since the vector $\psi(x_1,..,x_N)$ is selected a priori, condition (14) can easily be satisfied. The parameters c^* cannot be computed directly, however, because the second term in equation (13) is not known. They can be estimated sequentially by the following stochastic approximation algorithm, using discrete measurements in time n=1,2,... of the appropriate variables [1,2,3].

$$c_{n+1} = c_n + \gamma_n \left[\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \left(v_n(x_1^{i_1},..,x_N^{i_N}) - c_n^T \psi(x_1^{i_1},..,x_N^{i_N}) \right) \psi(x_1^{i_1},..x_N^{i_N}) \right] \quad (17)$$

The sequence of real numbers γ_n satisfies

$$\lim_{n\to\infty} \gamma_n = 0; \quad \sum_{n=1}^{\infty} \gamma_n = \infty; \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty \text{ and } E\{\|c_o\|\} < \infty \quad (18)$$

It is assumed that at the nth experiment the nI random observations of the

response equation (8) are independent, where I is given by equation (16).

It is shown in [1,7] that for algorithm (17) the estimate of the parameters converges in mean-square and with probability one to the parameters c^*, which minimize the performance criterion (11). Convergence in mean-square and with probability one are defined, respectively, as follows:

$$\lim_{n\to\infty} E\left\{\|c_n - c^*\|^2\right\} = 0 \tag{19}$$

$$P\left\{\lim_{n\to\infty} c_n = c^*\right\} = 1 \tag{20}$$

SENSITIVITY ANALYSIS

The subject of this section is the possibility of accelerating the convergence of algorithm (17) by choosing a priori the location of the measurement points $(x_1^{i_1},..,x_N^{i_N})$; $i_1=1,..,I_1;..;i_N=1,..,I_N$, such that a sensitivity criterion to be defined subsequently is maximized. Assume that the actual response can be expressed as follows:

$$u(x_1,..,x_N) = \sum_{k=1}^{K} c_k^* \psi_k(x_1,..,x_N)+e(x_1,..,x_N)=(c^*)^T\psi(x_1,..,x_N)+e(x_1,..,x_N) \tag{21}$$

If it can be assumed that the error term $e(x_1,..,x_N)$ is small in comparison to the first term in the expression above, and therefore it can be neglected, the performance criterion (11) becomes:

$$J(c) = \frac{1}{2} E\left\{ \sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\left[(c^*-c)^T\psi(x_1^{i_1},..,x_N^{i_N})+\eta(x_1^{i_1},..,x_N^{i_N})\right]^2\right\} \tag{22}$$

In view of conditions (9), (10), and the fact that the measurement points are fixed, the above relation yields:

$$J(c) = \frac{1}{2}(c-c^*)^T\left[\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\psi(x_1^{i_1},..,x_N^{i_N})\psi^T(x_1^{i_1},..,x_N^{i_N})\right](c-c^*) + \frac{1}{2}\sigma^2$$

$$= \frac{1}{2}(c-c^*)^T S(X_i)(c-c^*) + \frac{1}{2}\sigma^2 \tag{23}$$

The matrix $S(X_i)$ is defined by equation (14) and is positive definite. Minimizing the above expression with respect to the parameter vector c is equivalent to minimizing its first term only.

The fact that the matrix $S(X_i)$ equation (14) is positive definite can be used to define a number of sensitivity criteria. Among them are the following:

(a) Maximization of the minimum eigenvalue of $S(X_i)$ with respect to the measurement points.

(b) Maximization of the product of the eigenvalue of $S(X_i)$ with respect to the measurement points.

(c) Maximization of the trace of $S(X_i)$ with respect to the measurement points.

This list is not intended to be exhaustive. In this paper, maximization of the minimum eigenvalue $\lambda_{min}[S(X_i)]$ of $S(X_i)$ is chosen as the sensitivity criterion. Therefore, the measurement points will be chosen

such that $\lambda_{min}[S(X)]$ is maximized. However, in general the minimum eigenvalue is a multimodal surface in the measurement points. Consequently, the following random optimization algorithm [4,5] may be used to obtain the measurement points that maximize the minimum eigenvalue $\lambda_{min}[S(X)]$ readily computed from (14). The algorithm is:

$$\tilde{\alpha}_{n+1} = \begin{cases} \tilde{\alpha}_n & \text{if } \lambda_{min}[S(\tilde{\rho}_n)] - \lambda_{min}[S(\tilde{\alpha}_n)] \leq 0 \\ \tilde{\rho}_n & \text{if } \lambda_{min}[S(\tilde{\rho}_n)] - \lambda_{min}[S(\tilde{\alpha}_n)] > 0 \end{cases} \tag{24}$$

The vectors $\tilde{\alpha}$ and $\tilde{\rho}$ are defined as follows:

$$\tilde{\alpha}^T \triangleq \left[x_1^1,..,x_1^{I_1},\ x_2^1,..,x_2^{I_2},..,x_N^1,..,x_N^{I_N}\right] \tag{25}$$

$$\tilde{\rho}^T \triangleq \left[\tilde{\xi}_1^1,..,\tilde{\xi}_1^{I_1},\ \tilde{\xi}_2^1,..,\tilde{\xi}_2^{I_2},..,\tilde{\xi}_N^1,..,\tilde{\xi}_N^{I_N}\right] \text{ a random vector} \tag{26}$$

Define

$$I_k \triangleq I_1 + I_2 + \cdots + I_N \tag{27}$$

The I_k-dimensional parameter vector $\tilde{\alpha}$ is defined in the closed and bounded parameter space $\Omega_{\tilde{\alpha}}$ contained in Euclidean I_k-space. The space $\Omega_{\tilde{\alpha}}$ is defined as follows:

$$\Omega_{\tilde{\alpha}} = \left[0 < \tilde{\alpha}_1 < \frac{\pi}{I_1},..,\frac{\pi(I_1-1)}{I_1} < \tilde{\alpha}_{I_1} < \pi,..,0 < \tilde{\alpha}_{I_1+1} < \frac{\pi}{I_2},...,\frac{\pi(I_2-1)}{I_2} <\right.$$

$$\left.\tilde{\alpha}_{I_1+I_2} < \pi,..,0\ \pi\ \tilde{\alpha}_{I_1+I_2..+I_{N-1}+1} < \frac{\pi}{I_N},..,\frac{\pi(I_N-1)}{I_2} < \tilde{\alpha}_{I_1+..+I_N} < \pi\right] \tag{28}$$

In the definition above use has been made of the fact that the vector $X \in \Omega$, where Ω is defined by equation (3). The random variable $\tilde{\rho}$ is also defined in $\Omega_{\tilde{\alpha}}$ and its probability density satisfies $f_1(\tilde{\rho}) \neq 0$ for all $\tilde{\rho} \in \Omega_{\tilde{\alpha}}$. It may be decided that the points in each coordinate be uniformly spaced. Then the random variable $\tilde{\rho}$ equation (26) can be defined as follows:

$$\tilde{\rho} \triangleq \left[\tilde{\xi}_1^1,\ \tilde{\xi}_1^1 + \frac{\pi}{I_1},..,\tilde{\xi}_1^1 + \frac{\pi(I_1-1)}{I_1},\ \tilde{\xi}_2^1,\ \tilde{\xi}_2^1 + \frac{\pi}{I_2},..,\tilde{\xi}_2^1 + \frac{\pi(I_2-1)}{I_2},\right.$$

$$\left.\ldots,\ \tilde{\xi}_N^1,\ \tilde{\xi}_N^1 + \frac{\pi}{I_N},..,\tilde{\xi}_N^1 + \frac{\pi(I_N-1)}{I_N}\right] \tag{29}$$

where the random variable $\tilde{\xi}$ is reduced for this case to:

$$\tilde{\xi}^T = \left[\tilde{\xi}_1^1,\ \tilde{\xi}_2^1,\ldots,\tilde{\xi}_N^1\right] \tag{30}$$

The random variable $\tilde{\xi}$ is defined in the domain $\Omega_{\tilde{\xi}} \in E^N$ given by:

$$\Omega_{\tilde{\xi}} = \left[0 < \tilde{\xi}_1^1 < \frac{\pi}{I_1},\ 0 < \tilde{\xi}_2^1 < \frac{\pi}{I_2},\ldots,\ 0 < \tilde{\xi}_N^1 < \frac{\pi}{I_N}\right] \tag{31}$$

The above sensitivity analysis is illustrated in the following example where the acceleration of convergence is demonstrated.

EXAMPLE 1

Let the model of the system be the following:

$$a_1 \frac{\partial^2 u}{\partial x_1^2} + a_2 \frac{\partial^2 u}{\partial x_2^2} = 0; \quad x_1, x_2 \in \Omega \tag{32}$$

where u is a function of x_1 and x_2 and Ω is defined by:

$$\Omega \triangleq [0 < x_1 < \pi;\ 0 < x_2 < \pi] \tag{33}$$

The boundary conditions are:

$$u(x_1,0) = u(x_1,\pi) = u(0,x_2) = 0 \tag{34}$$

$$u(\pi,x_2) = g(x_2) \tag{35}$$

The function $g(x_2)$ is that plotted in Figure 1. The constant parameters a_1 and a_2 in (32) are assumed unknown. In the simulation used on the computer, a_1 and a_2 are assumed equal to unity. Therefore, (32) reduces to Laplace's equation:

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0 \tag{36}$$

The simulated solution of (36) with boundary conditions (34)-(35) obtained by the method of separation of variables [6] is:

$$u(x_1,x_2) = 8 \sum_{M=1,3,\text{odd}}^{3} \frac{(-1)^{\frac{M-1}{2}}}{M^2} \frac{\sin h(Mx_1)}{\sin h(M\pi)} \sin (Mx_2) \tag{37}$$

Harmonics greater than the third in the above solution are considered insignificant.

Let the solution be approximated by:

$$\hat{u}(x_1,x_2) = \sum_{m=1}^{3} c_m \psi_m(x_1,x_2) = c^T \psi(x_1,x_2) \tag{38}$$

The functions $\psi_m(x_1,x_2)$ are the following:

$$\psi_m(x_1,x_2) = \frac{\sin h(mx_1)}{\sin h(m\pi)} \sin(mx_2); \quad m = 1,2,3 \tag{39}$$

Three points in the x_1 coordinate and three points in the x_2 coordinate will be chosen for measurement [1]. Therefore, nine measurement points in Ω are used. That is, $(x_1^{i_1}, x_2^{i_2},)$; i_1=1,2,3; i_2=1,2,3 are chosen as the measurement points. From equation (38) the number of parameters to be estimated is three, and consequently, condition (15) is satisfied. Therefore the matrix $S(X_i)$ equation (14) in this case becomes:

$$S(X_i) = \sum_{i_1=1}^{3} \sum_{i_2=1}^{3} \psi(x_1^{i_1}, x_1^{i_2})\, \psi^T(x_1^{i_1}, x_2^{i_2}) \quad , \tag{40}$$

The function $\psi_m(x_1,x_2)$; m=1,2,3 are defined by equation (39). The minimum eigenvalue of $S(X_i)$, $\lambda_{min}[S(X_i)]$ computed by an appropriate subroutine, was maximized using algorithm (24) with the following definitions of the vector $\tilde{\alpha}$ equation (25) and $\tilde{\rho}$ equation (26)

$$\tilde{\alpha}^T \triangleq [x_1^1,\ x_1^2,\ x_1^3,\ x_2^1,\ x_2^2,\ x_2^3] \tag{41}$$

$$\tilde{\rho}^T \triangleq [\tilde{\xi}_1,\ \tilde{\xi}_2,\ \tilde{\xi}_3,\ \tilde{\xi}_4,\ \tilde{\xi}_4 + \frac{\pi}{3},\ \tilde{\xi}_4 + \frac{2\pi}{3}\ ] \tag{42}$$

Each of the random variables $\tilde{\xi}_1$, $\tilde{\xi}_2$ and $\tilde{\xi}_3$ is chosen from a uniform noise sequence over $[0,\pi]$. The random variable $\tilde{\xi}_4$ is chosen uniformly in $[0,\frac{\pi}{3}]$. From (42) the points in the x_1 direction are freely spaced while the points in the x_2 direction are uniformly spaced.

The minimum eigenvalue of S(X) was evaluated and its average value for twenty independent experiments is plotted in Figure 2. In order to emphasize the advantages of the sensitivity analysis, two cases were considered. The first one with almost arbitrary locations of the measurement points, the other one after the maximization of the λ_{min} For comaprison purposes values were properly arranged over 20 independent experiments.

Case 1:

The average values of the points for the twenty runs at the first iteration were:

$$\begin{aligned} x_1^1 &= 1.4092, & x_1^2 &= 1.4211, & x_1^3 &= 1.9814 \\ x_2^1 &= 0.4524, & x_2^2 &= 0.4524 + \frac{\pi}{3}, & x_2^3 &= 0.4524 + \frac{2\pi}{3} \end{aligned} \tag{43}$$

Case 2:

The average values of the points for the twenty runs at the 200th iteration were:

$$\begin{aligned} x_1^1 &= 2.6430, & x_1^2 &= 2.3018, & x_1^3 &= 2.8474 \\ x_2^1 &= 0.5218. & x_2^2 &= 0.5218 + \frac{\pi}{3}, & x_2^3 &= 0.5218 + \frac{2\pi}{3} \end{aligned} \tag{44}$$

Noisy measurements of the actual response (37) were assumed available as follows:

$$v(x_1^{i_1}, x_2^{i_2}) = u(x_1^{i_1}, x_2^{i_2}) + \eta(x_1^{i_1}, x_2^{i_2});\ i_1=1,2,3;\ i_2=1,2,3 \tag{45}$$

where the measurement noise satisfies conditions (9) and (10). From equation (37) the parameter vector c^* is defined by

$$c^{*T} \triangleq [8,0\ -\ \frac{8}{9}] \tag{46}$$

The unknown parameters c equation (38) were estimated using the stochastic approximation algorithm (17) and minimizing the performance criterion (11).

The sequence λ_n was chosen to be the following:

$$\lambda_n \triangleq \frac{1}{\lambda_{min}\left[n + \left(\frac{\lambda_{max}-\lambda_{min}}{2\lambda_{min}}\right)\right]} \tag{47}$$

The variables λ_{min} and λ_{max} are the minimum and maximum eigenvalues of $S(X_i)$ (41), respectively. Choosing this λ_n provides a common basis for comparing the convergence of algorithm (17) for Case 1 and Case 2. The samples of measurement noise were selected from a zero-mean Gaussian noise sequence. The example was programmed on the CDC 6500 digital computer. The results are ploted in Figure 3 for Case 1 and Case 2 and for two different variance of the measurement noise, namely, 0.01 and 1.0. The average parameter error for twenty independent experiments is plotted as a function of iteration. Figure 4 shows the standard deviation of the parameter error for the twenty experiments.

CONCLUSIONS

The method of stochastic approximations was successfully applied for the on-line estimation of a set of parameters in our approximate solution of systems described by linear elliptic partial differential equations with uncertainties in the mathematical model. Noisy measurements of the response were assumed available at a finite number of fixed points in the spatial domain. Sensitivity analysis was used to accelerate the algorithm. The minimum eigenvalue of a positive definite matrix obtained from the preselected approximating functions and the error criterion, which is a function of the measurement points was used as the sensitivity criterion. A random optimization algorithm was used to obtain the location of the measurement devices such that the sensitivity criterion was maximized. The simulation results for the example plotted in Figures 2-4 show that considerable acceleration of the algorithm is indeed possible by using the sensitivity analysis.

ACKNOWLEDGMENT

The authors wish to acknowledge Dr. G. Stein for his suggestions. This research was supported by NASA Institutional Grant NGR 15-005-021 and by Joint Services Electronics Program under ARPA Controact No. N00014-67-A-0226-0004.

REFERENCES

1. Badavas, P. C., "Control of a Class of Distribute Systems with Noisy Measurements", Ph.D. Thesis, Purdue University, August 1969.

2. Saridis, G. N., and Badavas, P. C., "Identification of Distributed Parameter Systems by Stochastic Approximation", Proceedings, 7th Symposium on Adaptive Processes, University of California, Berkeley, December 1968.

3. Kashyap, R. L., and Blaydon, C. C., "Recovery of Functions from Noisy Measurements Taken at Randomly Selected Points and Its Application in Pattern Recognition", Proc. of IEEE, Vol. 54, No. 8, August 1966, pp. 1127-29.

4. Matyas, J., "Random Optimization", Avtomatika i Telemekhanika, Vol. 26, No. 2 (Feb. 1966), pp. 246-53.

5. Gurin, L. S., "Random Search in the Presence of Noise", Eng. Cybernetics, No. 3 (May-June 1966), pp. 252-260.

6. Moon, P., and Spencer, D. E., Field Theory for Engineers, Princeton, New Jersey: D. Van Nostrand Company, Ind., 1961.

7. Saridis, G. N., and Badavas, P. C., "Control of a Class of Distributed Systems with Noisy Measurements", Purdue University, School of Electrical Engineering, Rept. No. TR-EE 69-9, 1969.

8. Saridis, G. N., Nicolic, Z. J., and Fu, K. S., "Stochastic Approximation Algorithms for System Identification, Estimation, and Decomposition of Mixtures", IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-5, No. 1 (Jan. 1969), pp. 8-15.

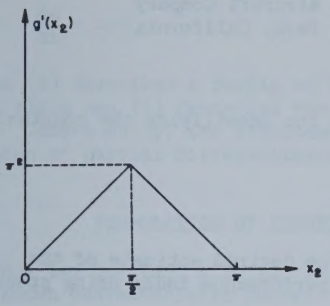


FIGURE 1. BOUNDARY CONDITION.

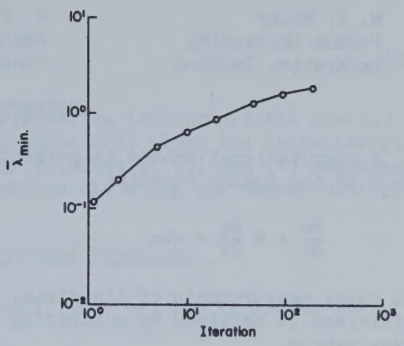


FIGURE 2. AVERAGE MINIMUM EIGENVALUE FOR EXAMPLE

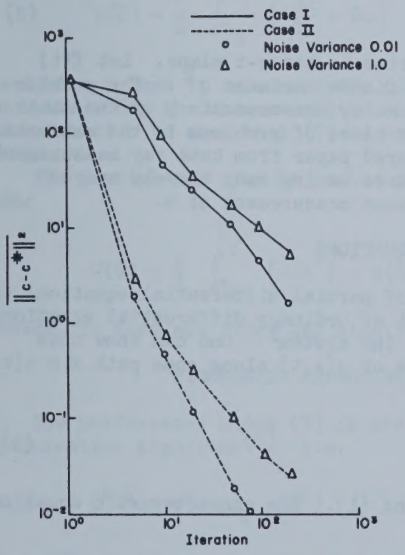


FIGURE 3. AVERAGE PARAMETER ERROR FOR EXAMPLE 1.

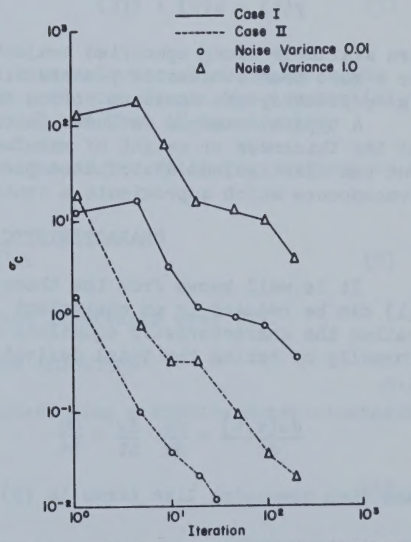


FIGURE 4. STANDARD DEVIATION OF THE PARAMETER ERROR FOR EXAMPLE

A RESULT ON THE IDENTIFICATION OF PARAMETERS IN DISTRIBUTED SYSTEMS REPRESENTED BY HYPERBOLIC EQUATIONS*

M. J. Wozny
Purdue University
Lafayette, Indiana

W. T. Carpenter
Hughes Aircraft Company
Conoga Park, California

ABSTRACT

A computational scheme is presented for identifying the constants a or b in the scalar equation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = -bu$$

from noisy measurements of the state. The desired estimate of the coefficient is obtained by minimizing a performance index using stochastic approximation.

STATEMENT OF PROBLEM

This paper is addressed to the following problem. Consider a distributed system described by the scalar partial differential equation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = -bu \tag{1}$$

where the constant a is known and the constant b is unknown. Assume that continuous noisy measurements of u, i.e.,

$$y(t) = u(t) + \xi(t) \tag{2}$$

are available along specified trajectories in the x-t plane. Let $\xi(t)$ be a zero mean stochastic process with finite variance σ^2. The problem is to identify the constant b from the noisy measurements y of the state u.

A typical example included in this class of problems is the estimation of the thickness or weight of manufactured paper from beta ray measurements. One can also include distributed processes having many closely spaced transducers which approximate a continuous measurement of u.

CHARACTERISTIC EQUATIONS

It is well known from the theory of partial differential equations that (1) can be reduced to an equivalent set of ordinary differential equations called the characteristic equations of the system[1]. One can show this formally by taking the total derivative of u(x,t) along some path x = x(t), i.e.

$$\frac{du(x,t)}{dt} = \frac{\partial u}{\partial x} \frac{dx}{dt} + \frac{\partial u}{\partial t} \tag{3}$$

and then comparing like terms in (3) and (1). The characteristic equations

*This work was supported by Joint Services Program, Contract Number N 00014-67-A0226-0004.

in this case are

$$\frac{dx}{dt} = a \tag{4}$$

$$\frac{du}{dt} = -bu \tag{5}$$

Equation (4) describes a family of trajectories (characteristic curves) in the x-t plane and (5) describes the solution u(t) along the characteristic curves. Consequently, the simultaneous solution of (3) and (4) represents a solution of partial differential equation (1) along the characteristic curves.

FORMULATION OF IDENTIFICATION PROCEDURE

Since the constant a in (1) is given, then (4) can be solved to obtain known characteristic curves

$$x - x_o = a(t - t_o) \tag{6}$$

where $x_o = x(t_o)$ represents the initial point on the curve. Consequently, the true value of b forces the solution of (5), along a known characteristic curve (6) to agree with the measured values of u along the same characteristic curve. This condition can be expressed in terms of a performance criterion as follows.

Let $\hat{b}$ denote an estimate of b and b^* the true value. Then b^* is characterized as that value of $\hat{b}$ which minimizes the performance index

$$J(\hat{b}) = \frac{1}{2} \int_{t_o}^{t_f} \left(\frac{du}{dt} + \hat{b}u \right)^2 dt \tag{7}$$

The limits t_o and t_f represent the time interval over which measurements are taken along the characteristic curve. One can invoke the exponential character of u to show that (7) is unimodal and that it has a unique minimum.

The problem can also be stated in terms of the following performance index

$$J(\hat{b}) = \frac{1}{2} \int_{t_o}^{t_f} [u(b^*) - u(\hat{b})]^2 dt \tag{8}$$

However, it does not offer any significant advantages over (7).

STOCHASTIC APPROXIMATION ALGORITHM

The performance index (7) is minimized using a Robbins-Munro stochastic approximation algorithm[2], i.e.

$$b^{k+1} = b^k - \rho_k \, m(\hat{b}^k) \tag{9}$$

where $m(\hat{b})$ is the regression function of the gradient of $J(\hat{b})$ and the sequence ρ_k is chosen to satisfy

$$\sum_{k=1}^{\infty} \rho_k = \infty \tag{10}$$

$$\sum_{k=1}^{\infty} \rho_k^2 = c < \infty \tag{11}$$

which guarantess convergence. Thus all that remains is to determine the required regression function.

REGRESSION FUNCTION

The gradient of $J(\hat{b})$ is

$$\frac{\partial J(\hat{b})}{\partial \hat{b}} = \int_{t_o}^{t_f} \left(\frac{du}{dt} + \hat{b}u\right) u \, dt \tag{12}$$

Assume that the noise introduced in the system is additive measurement noise as defined in (2). Differentiating (2) along a characteristic curve yields a noisy estimate of the derivative of u, i.e.,

$$\frac{dy}{dt} = \frac{du}{dt} + \frac{d\xi}{dt} \tag{13}$$

The regression function of (12) can be expressed as

$$m(\hat{b}) = E\left\{ \int_{t_o}^{t_f} \left[\left(\frac{dy}{dt} + \hat{b}y \right) y - B(\hat{b}) \right] dt \right\} \tag{14}$$

The bias term $B(\hat{b})$ is evaluated from

$$m(b^*) = 0 = E\left\{ \int_{t_o}^{t_f} \left[\left(\frac{d\xi}{dt} + b^*\xi \right) (u + \xi) - B(b^*) \right] dt \right\} \tag{15}$$

which reduces to

$$B(\hat{b}) = \hat{b}\,\sigma^2 + \left. \frac{dR_{\xi\xi}(\tau)}{d\tau} \right|_{\tau = 0} \tag{16}$$

If the autocorrelation function of the transducer noise is differentiable at $\tau = 0$, then the second term in expression (16) is zero. If the noise process is Gaussian, one can apply the property that a Gaussian process and its derivative are uncorrelated. In the more general case one would resort to state estimation to find u and then differentiate this signal to obtain the derivative of u directly.

Summarizing, the computational procedure first requires that one find a characteristic curve and measure u along this curve. Next compute $\frac{du}{dt}$ from this measured data. Then guess a value of $b = \hat{b}^o$ and compute successive improvements b^{k+1} using (9) where $m(\hat{b})$ is defined by (14) and (16).

EXAMPLE

The following example was programmed on a digital computer. The system to be identified was simulated on the computer from given initial and boundary value data. Measurements were made on the solutions and noise, generated by a digital pseudo-random number generator program, added to simulate measurement errors. The algorithm discussed then processed the noisy measurements and extracted the estimates of the unknown parameters.

The system used in the example was

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = -bu \tag{17}$$

where the true value of b is 1. The input was a unit ramp and the integral along the characteristic was approximated by trapezoidal integration with a base of .01. The results are shown in Figure 1, where the initial gain in the algorithm was $\rho_o = .01$ and the initial guess of b was $\hat{b}^o = .5$.

CONCLUSION

An algorithm for the identification of the constant coefficient b from noisy measurements was presented. A similar algorithm for the identification of a can be developed. The scheme has the advantage that it converges regardless of the noise statistics but the disadvantage that bias terms occur and must be estimated. The performance index in (8) also leads to bias terms involving autocorrelation functions.

There are many directions in which these ideas can be expanded. The most obvious extension is to the identification of nonlinear coefficients in a system of first order hyperbolic partial differential equations.

Another important class of problems is identification using limited measurement points. This idea can also be extended to more practical distributed models which involve diffusion type terms. Some of these topics are discussed in reference 3.

REFERENCES

1. Courant, R., and Hilbert, D., Methods of Mathematical Physics, Vol. 2, Interscience Publishers, New York, 1966.

2. Wilde, D. J., Optimum Seeking Methods, Prentic Hall, Inc., Englewood Cliffs, New Jersey, 1964.

3. Carpenter, W., Wozny, M., and Goodson, R., "Distributed Parameter Identification Using the Method of Characteristics," to be published.

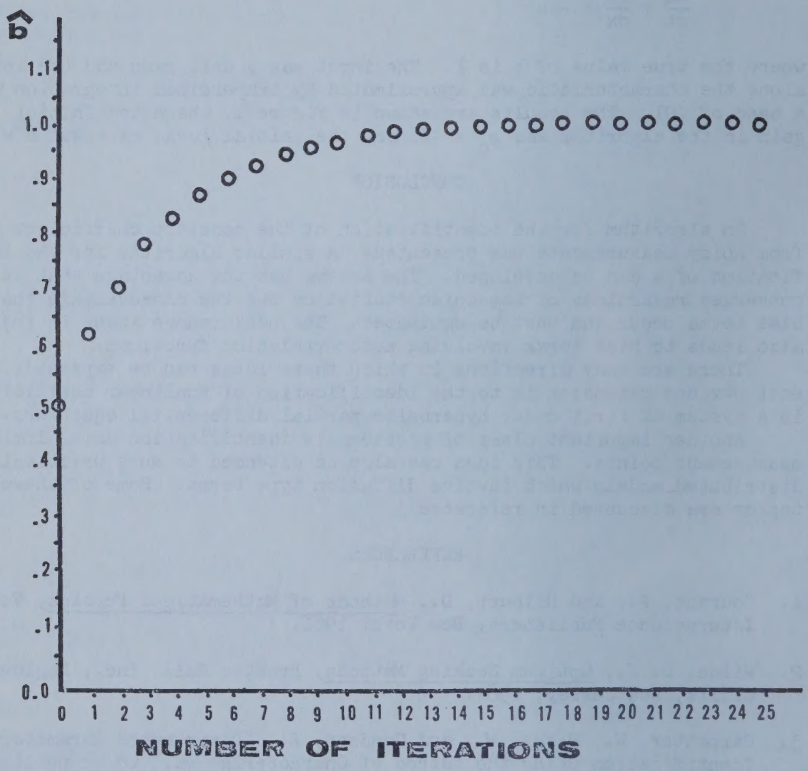


Figure 1. Identification of b

ESTIMATION AND IDENTIFICATION OF NONLINEAR SYSTEM WITH CONSTRAINTS

D. ISAACS
Associate Professor
University of California, Irvine
Irvine, California

W. LOCHRIE
North American Rockwell Fellow
University of California, Irvine
Irvine, California

ABSTRACT

A recursive algorithm is developed for the estimation of the states and identification of the parameters of a system with nonlinear dynamics in which linear constraints are imposed on these states and parameters. A quadratic programming computational method is used in this algorithm. An example involving the estimation of the trajectory of a reentry vehicle is presented in detail.

I. INTRODUCTION

In many estimation and identification problems, physical or mathematical reasoning imposes constraints on the variables being computed. The quality of the estimates can obviously be improved if this constraint information is incorporated in the computational algorithm employed. In a previous work [1], the authors developed a recursive estimation algorithm for systems with linear dynamics in which constraints were imposed on the state variables. In an extension of that work, this paper considers the estimation of the states and identification of the parameters of systems with nonlinear dynamics in which linear constraints are imposed on the state trajectory and the parameter values. The recursive computational algorithm developed utilizes a quadratic programming technique such as Wolfe's method [2]. In the limit of imposing no constraints the algorithm reduces to the Kalman filter for nonlinear systems [3,4].

II. PROBLEM FORMULATION

Consider the nonlinear dynamic system

$$\dot{\underline{\zeta}} = \underline{f}(\underline{\zeta}, \underline{\pi}) \tag{1}$$

where $\underline{\zeta}(t)$ is the state vector and $\underline{\pi}$ is the parameter vector. Assume it is desired to estimate the trajectory $\underline{\zeta}(t)$ and identify the vector of constants $\underline{\pi}$. Also assume that the constraints on the state trajectory

and the parameter values are either linear or can be approximated by linear constraints of the form

$$A\,\underline{\zeta} \leq \underline{\alpha} \tag{2}$$

$$B\,\underline{\pi} \leq \underline{\beta} \tag{3}$$

Let observations be made at $t_1, t_2, \ldots, t_k, \ldots$ and assume that they are related to the states and parameters by

$$\underline{Y}_k = \underline{g}(\underline{\zeta}_k, \underline{\pi}) + \underline{v}_k \tag{4}$$

For notational convenience the subscript k on matricies and vectors denotes evaluation at time t_k. The vector $\underline{v}_k$ is zero mean white measurement noise with covariance denoted by Q_k^{-1}.

III. COMPUTATIONAL ALGORITHM

System Description

The computational algorithm is obtained by first assuming that the estimation error can be described by a linear expansion about the estimated trajectory $\tilde{\underline{\zeta}}(t)$ and estimated parameter vector $\tilde{\underline{\pi}}$. The algorithm developed by the authors [1] is then modified to handle the resulting linearized equations.

Let the current estimates of $\underline{\zeta}_k$ and $\underline{\pi}$ be denoted by $\tilde{\underline{\zeta}}_k$ and $\tilde{\underline{\pi}}_k$ respectively. The estimate of $\underline{\zeta}$ is propagated by the dynamical equation

$$\dot{\tilde{\underline{\zeta}}} = \underline{f}(\tilde{\underline{\zeta}}, \tilde{\underline{\pi}}) \tag{5}$$

Let the error in the knowledge of $\underline{\zeta}$ and $\underline{\pi}$ be represented by $\underline{z}$ and $\underline{p}$ respectively. That is,

$$\underline{z}_k = \underline{\zeta}_k - \tilde{\underline{\zeta}}_k \text{ and } \underline{p}_k = \underline{\pi} - \tilde{\underline{\pi}}_k \tag{6}$$

Expanding (1) about the estimates $\tilde{\underline{\zeta}}$ and $\tilde{\underline{\pi}}$ and keeping only the first order terms yields the following linear differential equation describing the errors

$$\dot{\underline{z}}(t) = F(t)\,\underline{z}(t) + G(t)\,\underline{p} \tag{7}$$

F(t) and G(t) are partial derivative matrices $\frac{\partial \underline{f}}{\partial \underline{\zeta}}$ and $\frac{\partial \underline{f}}{\partial \underline{\pi}}$ respectively.

The notation $\frac{\partial \underline{f}}{\partial \underline{\zeta}}$ denotes a matrix of partial derivatives where the ij<u>th</u>

element is $\frac{\partial f_i}{\partial \zeta_j}$. The partial derivatives in (7) are evaluated along the assumed solution $\underset{\sim}{\zeta}(t)$.

The solution to (7) is given by

$$\underline{z}(t) = \Theta(t,t_k)\ \underline{z}(t_k) + \Lambda(t,t_k)\ \underline{p} \tag{8}$$

where $\Theta(t,t_k)$ is the transition matrix and $\Lambda(t,t_k)$ is the sensitivity matrix of system states to parameter errors. They are obtained by integration of the following matrix differential equations.

$$\dot{\Theta}(t,t_k) = F(t)\ \Theta(t,t_k) \quad ;$$

$$\Theta(t_k,t_k) = I \quad \text{(Identity Matrix)} \tag{9}$$

and

$$\dot{\Lambda}(t,t_k) = F(t)\ \Lambda(t,t_k) + G(t) \ ;$$

$$\Lambda(t_k,t_k) = 0 \ \text{(Null Matrix)} \tag{10}$$

This solution can be verified by simply differentiating (8) and using (9) and (10).

Since it is desired to get the equations in a format such that the algorithm in [1] can be implemented make the following definitions:

$$\underline{x}(t) \triangleq \begin{bmatrix} \underline{z}(t) \\ \\ \underline{p} \end{bmatrix}; \quad \Theta_k \triangleq \Theta(t_{k+1}, t_k); \ \Lambda_k \triangleq \Lambda(t_{k+1},t_k) \tag{11}$$

and

$$\Phi_k \triangleq \begin{bmatrix} \Theta_k & \Lambda_k \\ \\ 0 & I \end{bmatrix} \tag{12}$$

Thus (8) can now be written simply as

$$\underline{x}_{k+1} = \Phi_k\ \underline{x}_k \tag{13}$$

The transition matrix of the adjoint system, $\Psi_k \triangleq (\Phi_k^{-1})^T$, is also

necessary for the algorithm in [1]. For notational convenience introduce the matrix $\Omega_k \triangleq (\Theta_k^{-1})^T$. It can be easily verified that the matrix Ψ_k is given by:

$$\Psi_k = \begin{bmatrix} \Omega_k & 0 \\ -\Lambda^T{}_k \, \Omega_k & I \end{bmatrix} \tag{14}$$

where Ω_k is computed by integrating from t_k to t_{k+1} the differential equation

$$\dot{\Omega}(t,t_k) = - F^T(t) \, \Omega(t,t_k)$$

with

$$\Omega(t_k, t_k) = I \tag{15}$$

The matrix Λ_k is computed by integrating (10) from t_k to t_{k+1}.

Measurement Process

Now consider the process of making a measurement. The vector of actual measurements $\underline{\gamma}_k$ is given by (4). Denote the expected measurements at t_k by

$$\tilde{\underline{\gamma}}_k = \underline{g}(\tilde{\underline{\zeta}}_k, \tilde{\underline{\pi}}_k) \tag{16}$$

Let $\underline{y}_k$ denote the measurement residual

$$\underline{y}_k = \underline{\gamma}_k - \tilde{\underline{\gamma}}_k \tag{17}$$

This residual is related to the state vector error and parameter vector error, to the first order, as

$$\underline{y}_k = J_k \, \underline{z}_k + K_k \, \underline{p}_k + \underline{v}_k \tag{18}$$

where $$J_k = \frac{\partial \underline{g}}{\partial \underline{\zeta}} (t_k) \quad \text{and} \quad K_k = \frac{\partial \underline{g}}{\partial \underline{\pi}} (t_k) \tag{19}$$

Again it is desirable to get the equations in a format compatible with the notation in [1]. Hence, introduce another matrix

$$M_k \triangleq \begin{bmatrix} J_k & K_k \end{bmatrix} \tag{20}$$

Thus (18) can be expressed as

$$\underline{y}_k = M_k \underline{x}_k + \underline{v}_k \tag{21}$$

Imposition of Constraints

Next consider the imposition of the constraints given in (2) and (3). First, note that the state and parameter vectors are updated in the following fashion when the measurement at time t_k is processed:

$$\tilde{\underline{\zeta}}_{k+} = \tilde{\underline{\zeta}}_{k-} + \hat{\underline{z}}_k \tag{22}$$

and

$$\tilde{\underline{\pi}}_{k+} = \tilde{\underline{\pi}}_{(k-1)+} + \hat{\underline{p}}_k \tag{23}$$

The state trajectory prior to correction, $\tilde{\underline{\zeta}}_{k-}$, is obtained by integration of (5) with initial condition $\tilde{\underline{\zeta}}_{(k-1)+}$. Note that $\tilde{\underline{\zeta}}_{0+}$ and $\tilde{\underline{\pi}}_{0+}$ are simply the initial estimates of the state and parameter vectors respectively. The symbols $\hat{\underline{z}}_k$ and $\hat{\underline{p}}_k$ denote the estimates of the error vectors $\underline{z}_k$ and $\underline{p}$ respectively. These estimates are chosen to minimize the mean square estimation error subject to the constraints:

$$A_k \underline{\zeta}_{k+} \leq \underline{\alpha} \tag{24}$$

$$B \underline{\pi}_{k+} \leq \underline{\beta} \tag{25}$$

In order to make the technique computationally tractable the constraints are imposed on the trajectory only at time t_k rather than for all admissible times.

Since the estimation process involves $\underline{z}_k$ subtract $A\underline{\zeta}_{k-}$ from both sides of (24)

$$A_k \underline{z}_k \leq \underline{\alpha} - A \underline{\zeta}_{k-} \tag{26}$$

In a similar fashion (25) becomes

$$B \underline{p}_k \leq \underline{\beta} - B\underline{\pi}_{k-} \tag{27}$$

Combining (26) and (27) into one matrix equation yields

$$H_k \underline{x}_k \leq \underline{d}_k, \tag{28}$$

where

$$H_k = \begin{bmatrix} A_k & 0 \\ 0 & B \end{bmatrix} \quad \text{and} \quad \underline{d}_k = \begin{bmatrix} \underline{\alpha} - A\underline{\zeta}_{k-} \\ \underline{\beta} - B\underline{\pi}_{k-} \end{bmatrix} \tag{29}$$

The algorithm for nonlinear systems can now be obtained from the linear system algorithm [1]. Note that since the errors are being estimated $\underline{x}_0 = \underline{0}$, which thus implies $\underline{c}_0 = \underline{0}$. Therefore, since the state and parameter vectors are re-initialized at each measurement the $\psi_k \underline{c}_k$ term is not needed in computing $\underline{c}_{k+1}$. The algorithm is as follows:

(i) Compute ψ_k given in (14). This requires integrating equations (10) and (15) from t_k to t_{k+1}.

(ii) Update the cost functional $f(\hat{\underline{x}}_{k+1})$

$$\begin{aligned} &(1)\quad D_{k+1} = \psi_k D_k \psi_k^T + M_{k+1}^T Q_{k+1} M_{k+1} \\ &\qquad\qquad \text{with } D_0 = \text{cov}(\underline{x}_0)^{-1} \\ &(2)\quad \underline{c}_{k+1} = 2 M_{k+1}^T Q_{k+1} \underline{y}_{k+1} \\ &(3)\quad f(\hat{x}_{k+1}) = -\hat{x}_{k+1}^T D_{k+1} \hat{\underline{x}}_{k+1} + \underline{c}_{k+1}^T \hat{\underline{x}}_{k+1} \end{aligned} \tag{30}$$

(iii) Solve the quadratic programming problem

$$\text{Maximize } f(\hat{x}_{k+1}) \quad \text{subject to} \quad H_{k+1} \hat{\underline{x}}_{k+1} \leq \underline{d}_{k+1}$$

(iv) Update the state vector and parameter vector as shown in (22) and (23).

IV REENTRY EXAMPLE

The problem of estimating the trajectory of a reentry vehicle (RV) was solved using the algorithm given above. It was assumed that knowledge of the target area and vehicle manuvering capability made it possible to impose constraints on the trajectory. A two dimensional model was used with the coordinate system defined such that, at $t = 0$, ζ_1 was radial and ζ_2 was downrange with respect to the RV. The equations of motion of the RV are as follows:

$$
\begin{aligned}
\dot{\zeta}_1 &= \zeta_3 \\
\dot{\zeta}_2 &= \zeta_4 \\
\dot{\zeta}_3 &= -\frac{\mu}{R^3}\zeta_1 - \gamma V \rho_0 e^{-\beta h}\zeta_3 \\
\dot{\zeta}_4 &= -\frac{\mu}{R^3}\zeta_2 - \gamma V \rho_0 e^{-\beta h}\zeta_4
\end{aligned}
\tag{31}
$$

where R is the magnitude of the RV position vector and V is the magnitude of the velocity vector. The altitude, h, is simply $R-R_e$ where R_e is the radius of the earth (20,900,000 feet). The gravitational constant, μ, has the value 1.4076452×10^{16}. The vehicle drag coefficient γ is assumed to be constant at 1.6 $ft^2/slug$. The atmospheric density parameters ρ_0 and β are given by .0027 $slug/ft^3$ and 4.25×10^{-5} ft^{-1} respectively [5]. The computation of the matrix of partial derivatives, F, is straight forward and therefore will not be given explicitly.

It is assumed that the RV is tracked by radar which measures the range and direction of the RV with respect to the radar site. The range, denoted by d, is given as follows;

$$
d = [\,(\zeta_1 - r_1)^2 + (\zeta_2 - r_2)^2\,]^{1/2} \tag{32}
$$

where (r_1, r_2) denotes the radar site position vector. The gimbal angle measurement is given by the equation

$$
\gamma = \arcsin \frac{\zeta_1 - r_1}{d} \tag{33}
$$

The output matrix relating the state estimation error to the measurement residual is given by the equation:

$$
M_k = \begin{bmatrix}
\frac{\zeta_1 - r_1}{d} & \frac{\zeta_2 - r_2}{d} & 0 & 0 \\
\frac{\zeta_2 - r_2}{d^2} & \frac{-(\zeta_1 - r_1)}{d^2} & 0 & 0
\end{bmatrix} \tag{34}
$$

where the elements of the matrix are evaluated at time t_k. In the simulation the radar station was assumed to have coordinates $(R_e, 0)$. It

was assumed that the penetration aids of the RV cause high values of observation noise. An rms range error of 30,000 feet and rms angular error of .05 radians was used. Two sets of constraints were used as shown in Figure 1. The constraints were specified as follows:

$$A = \begin{bmatrix} 1 & S_{11} & 0 & 0 \\ 1 & S_{21} & 0 & 0 \end{bmatrix} \quad ; \underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \tag{35}$$

When $\zeta_2 <$ 175,000 feet: $S_{11} = .7$; $S_{21} = -.07$; α_1 - 21,325,000 and $\alpha_2 = -21{,}175{,}000$. For $\zeta_2 >$ 175,000 feet: $S_{11} = .4$, $S_{21} = -.4$; $\alpha_1 =$ 21,270,000 and $\alpha_2 =$ - 21,230,000.

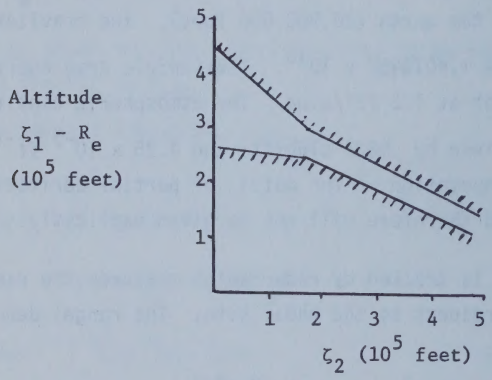


Fig. 1 Constraints for Example

The initial rms position error was assumed to be 75,000 ft per channel and initial rms velocity error 75 feet/sec per channel. A set of measurements was made every 5 seconds for 25 seconds. The results of the Monte Carlo simulation are shown in Fig. 2. One Monte Carlo run consisting of 5 measurements for both Kalman and the constrained estimator took approximately 2 seconds on the IBM 360/91.

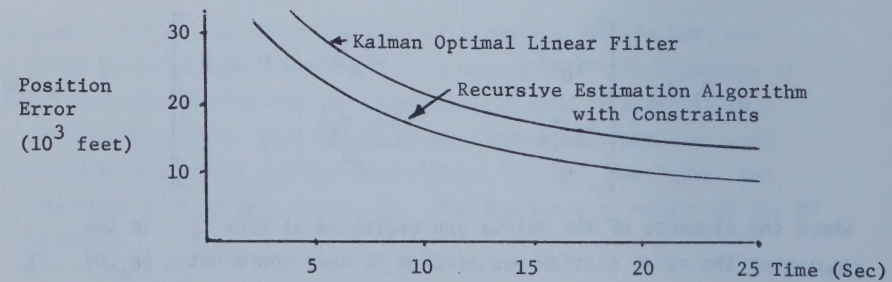


Fig. 2 Results of Monte Carlo Simulation

CONCLUSIONS

A method has been presented which permits utilization of constraint information in the estimation and identification of nonlinear systems. The method was shown to provide a computationally feasible method of solution in a relatively complex example. It is felt that this technique has practical application in systems where state variable constraints and measurement noise are such that the constraints are activated frequently. The numerical example in which the constraints were activated often demonstrates the improvement in performance that is possible.

REFERENCES

1. Isaacs, D., Lochrie, W.D., "Nonlinear Programming and Recursive Optimal Estimation," Proceedings of the Sixth Annual Allerton Conference on Circuit and System Theory, October, 1968.

2. Wolfe, P., "The Simplex Method for Quadratic Programming," Econometrica, 27, 1959, pp.382-398.

3. Smith, G.L., Schmidt, S.F., and McGee, L.A., "Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity,: NASA TR-R-135 1962.

4. Gunckel, T.L., "Orbit Determination Using Kalman's Method," Vol. 10 No. 3, pp. 273-291, Journal of the Institute of Navigation, Autumn, 1963.

5. W.H.T. Loh, Dynamics and Thermodynamics of Planetary Entry, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963, pp.12-14.

MAJORITY LOGIC DECODING FOR LINEAR CODES

PAUL LAITY
Control Data Corporation
St. Paul, Minnesota

K. C. WANG
Washington State University
Pullman , Washington

Abstract--This paper presents a generalized approach to the decoding of linear error-correcting codes. Properties of the generating matrix are used to define several decoding schemes which can be applied to any linear code.

PRELIMINARIES

Let V^k be the vector space of k-tuples and V^n be the vector space of n-tuples over the finite field GF(q). Let C be any (n, k) linear code with an n x k generating matrix $M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{k1} & \cdots & & a_{kn} \end{pmatrix}$ $a_{ij} \in GF(q)$

Then the row space of $M = C \subset V^n$.

Let α be a 1-to-1 mapping of V^k onto C defined as follows:

$$\text{for } c = (c_1, \ldots, c_k) \in V^k$$

$$\alpha c = (b_1, \ldots, b_n) \quad \text{where}$$

$$b_j = \sum_{i=1}^{k} a_{ij} c_i$$

Thus, if the mapping α is used to encode a set of k-information symbols and the code word $b = (b_1, \ldots, b_n)$ is received, the information symbols $(c_1, \ldots, c_k)$ can be recovered by solving a subset of k linear equations taken from the following set of n equations in k unknowns.

$$b_1 = a_{11} c_1 + a_{21} c_2 + \ldots + a_{k1} c_k$$

$$b_2 = a_{12} c_1 + a_{22} c_2 + \ldots + a_{k2} c_k$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$b_n = a_{1n} c_1 + a_{2n} c_2 + \ldots + a_{kn} c_k$$

Each different subset of k linearly independent equations will give one solution for c. The number of such subsets can be determined by the following:

Theorem 1 C has a minimum distance equal to d iff. every set of n - d + 1 column vectors of m span V^k.

Proof: Suppose C has a minimum distance equal to d.
Let $b = (b_1, \ldots, b_n) \in C$. Then $\exists$ constants $c_i \in GF(q)$, $i = 1, \ldots, n$ $\ni$

$$b_j = \sum_{i=1}^{k} c_i a_{ij} \Rightarrow \exists \; c = (c_1, \ldots, c_k) \; V^k \ni (c \mid a_j) = b_j$$

where a_j is the j-th column vector of m and $(\ \mid\ )$ denotes the Hilbert space product defined on V^k as follows:

$$\text{for } x, y \in V^\kappa \quad (x|y) = \sum_{i=1}^{k} x_i y_i$$

Since C has a minimum distance equal to d, b has at least d non-zero components.

$\Rightarrow$ every set of n - d + 1 components of b contains at least one non-zero element.

$\Rightarrow \nexists$ a vector $c \in V^k \ni (c \mid a_j) = 0 \; \forall$ column vector in any set of n - d + 1 column vectors.

$\Rightarrow \forall$ set of n - d + 1 column vector $\nexists$ a vector orthogonal to every vector in the set.

$\Rightarrow$ every set of n - d + 1 column vectors span V^k.

Conversely let $\{a_j\}$ denote an arbitrary set of n - d + 1 column vectors; then $\{a_j\}$ spans V^k.

$\Rightarrow \nexists$ a vector orthogonal to every vector in $\{a_j\}$.

$\Rightarrow \nexists \; c \in V^k \ni (c|a_j) = 0 \; \forall \; a_j \in \{a_j\}$

$\Rightarrow \nexists \; b \in C$ with more than d non-zero components.

$\Rightarrow$ C has a minimum distance equal to d.

It follows immediately from theorem 1 that the maximum minimum distance, $d_{max.}$, for any (n, k) linear code C is equal to n - k + 1 and that $d = d_{max.}$ iff. every set of k column vectors are linearly independent. (If $q = 2$, $d = d_{max.}$ only when $n - k = 1$).

DECODING PROCEDURES

I.

Every set of n - k + 1 column vectors define a set of n - d + 1 equations in k unknowns. Let each of these sets be denoted by s_i, $i = 1, \ldots, \binom{n}{n-d+1}$ and let e^i_j, $j = 1, \ldots, n - d + 1$ denote the n - d + 1 equations contained in s_i. Then $\{e^i_j \mid j = 1, \ldots, n - d + 1\} = \{e^i_j\} = s_i$.

Each $e_j^i \in s_i$ defines a hyperplane. If $\{e_j^i\}$ are all correct, we know that $n - d + 1$ hyperplanes defined by $\{e_j^i\}$ will intersect at a single point. If $\{e_j^i\}$ are not all correct, either the hyperplanes defined by $\{e_j^i\}$ will intersect at a single wrong point or more than one intersection point will be defined.

We will let each s_i which defines hyperplanes intersecting at a single point represent one vote (to be used in determining $c = (c_1, \ldots, c_k)$). Any s_i which defines hyperplanes intersecting at more than one point must contain wrong equations and thus will not be used to determine c. Thus for s errors we have $\binom{n-s}{n-d+1}$ correct votes and:

Lemma 1: For s errors it is possible to get at most $\binom{s+n-d}{n-d+1}$ wrong votes for c which are all alike.

Proof: It may happen that there does not exist a subset of s_i of order $(n - d + 1) - 1$ (denoted by s_i) which defines hyperplanes intersecting at a point and, $s_i \cup e_i^k$, where e_i^k is an incorrect equation, defines a single incorrect intersection point.

Similarly any combination of wrong equations when added to s_i may define a single wrong intersection point. Thus the largest number of hyperplanes which can intersect at a single wrong point is equal to $s + (n - d + 1) - 1 = s + n - d$ where s is the number of equations in error and the remaining equations are chosen from the $n - s$ correct ones. Any more correct hyperplanes would determine a correct solution; i.e., a different intersection point.

Thus if s errors are made, we have $\binom{n-s}{n-d+1}$ correct votes and at most $\binom{s+n-d}{n-d+1}$ incorrect votes which are all alike. To insure correct decoding the number of correct votes must exceed the number of alike incorrect votes, i.e.,

$$\binom{n-s}{n-d+1} > \binom{s+n-d}{n-d+1}$$

$$\Rightarrow \quad n - s > s + n - d$$

$$\Rightarrow \quad 2s < d$$

$$\Rightarrow \quad s \leq \frac{d}{2} - 1 \qquad \text{for d even}$$

$$s \leq \frac{d-1}{2} \qquad \text{for d odd}$$

We have shown that all linear codes are decodeable by a voting scheme and that the number of random errors which can be corrected is equal to the maximum capability of the code.

II.

Every set of $n - d + 1 + s$ equations contains a subset of $n - d + 1$ correct equations. Let s_i, $i = 1,\ldots, \binom{n}{n - d + 1 + s}$ represent the $\binom{n}{d - 1 - s}$ sets of $n - d + 1 + s$ equations. It can be determined if s_i contains any incorrect equations by checking the $\binom{n - d + 1 + s}{n - d + 1}$ solutions generated by s_i. If all the solutions derived from s_i are the same, s_i cannot contain any incorrect equations, and if s_i contains any incorrect equations the solutions cannot all be the same since one correct solution will always result from the $n - d + 1$ correct equations. Thus in order to decode correctly when s equations are in error we must have at least one s_i which contains all correct equations, i.e.,:

$$n - (n - d + 1 + s) \geq s$$

$$d - 1 \geq 2s$$

$$s \leq \frac{d - 1}{2} \qquad \text{for d odd}$$

and $$s \leq \frac{d}{2} - 1 \qquad \text{for d even}$$

Thus any linear code may be decoded by checking each of the $\binom{n}{s}$ different sets of $n - d + 1 + s$ equations.

III.

Since $n - s > n - d + 1$, any set of $n - s$ equations may be used to solve for c. If all combinations of s equations are removed from the set of n equations, one set must contain all correct equations and will thus result in a correct solution for c. The correct solution can be found from the set of $\binom{n}{s}$ solution sets by calculating the code word corresponding to each different solution and comparing it to the received code word. If $s \leq \frac{d - 1}{2}$, for d odd, or $s \leq \frac{d}{2} - 1$, for d even, the correct solution will be the only solution which generates a code word which is the same as the received code word (within s components).

Also any burst error which does not change the received code word to within an s component of a different code word will be corrected.

REFERENCES

1. W. W. Peterson, Error-Correcting Codes, New York: Wiley 1961

2. I. S. Reed and G. Solomon, "Polynominal Codes Over Certain Finite Fields," J. SIAM, vol. 8, pp. 300-304, 1960

3. T. Kasami, S. Lin, and W. W. Peterson, "New Generalizations Of The Reed-Muller Codes-Part I: Primitive Codes," IEE Trans. Information Theory, vol. IT-14, pp. 189-199, March 1968

4. T. Kasami, S. Lin, and W. W. Peterson, "Polynominal Codes" IEE Trans. Information Theory, vol. 14 No. 6 November, 1968

5. E. J. Weldon, Jr. "Some Results On Majority-Logic Decoding," Error Correcting Codes. Edited by Mann, New York Wiley 1968

6. Jeremy J. Stone, "Multiple-Burst Error Correction With The Chinese Remainder Theorem" J. SIAM Vol. 11, pp. 74-81, 1963

7. S. K. Berberian, Introduction to Hilbert Space. New York, Oxford University Press, 1961.

8. C. B. Hanneken, Introduction to Abstract Algebra. Belmont, Calif. Dickenson Publishing Co., 1968

A LARGE CLASS OF CYCLIC AND SHORTENED CYCLIC BINARY CODES FOR MULTIPLE ERROR CORRECTION

HARRY T. HSU
Department of Computer Science
Pennsylvania State University
University Park, Pennsylvania 16802

ABSTRACT

This paper is a study of a large class of cyclic and shortened cyclic binary codes for multiple error correction. The major result is: If $g(x) = \prod_{i=1}^{2\ell} g_i(x)$ generates a basic BCH code of the length N with minimum weight $\geq 2\ell t+1$, then $g(x) = \prod_{j=1}^{\ell} (g_j(x))^{i_j} \cdot \left(\prod_{j=(\ell+1)}^{2\ell} g_j(x) \right)$ where $i_1 > i_2 > - - - \geq i_\ell \geq 1$ generates a shortened cyclic or non-shortened cyclic code of the length $i_1 N$ with minimum weight at least $2\ell t+1$. Three classes of cyclic codes will be presented. Class I and Class III codes are cyclic codes. Class II codes are shortened cyclic codes.

I. INTRODUCTION

Since the discovery of the Bose-Chauhuri-Hocquenghem Codes (abbreviated as BCH codes) (1960), cyclic codes for multiple-error correction and their implementation have been the most widely treated subjects. This paper is a study of the non-BCH cyclic and shortened cyclic codes for multiple error correction. A large class of cyclic and shortened cyclic codes for multiple error correction will be presented. Class I codes consist of cyclic codes which have generator polynomials of the form $g(x) = (g_1(x))^{2^m} g_2(x)$. Class II codes consist of non-BCH-shortened-cyclic codes. Class III codes consist of non-BCH-cyclic codes. Part II of this paper will deal with preliminaries, part III deals with Class I codes, part IV deals with Class II codes, part V deals with Class III codes.

II. PRELIMINARIES

In order to make this paper relatively self-contained, some of the results which are already known will be briefly reviewed. For general reference, the reader is referred to Peterson (1961) and Berlekamp (1968). The following notations will be used in this paper:

α	an arbitrary element of $GF(2^f)$.
N	the order of α.
$g_1(x)$	has all the α^i, $1 \leq i \leq t$ as roots.
$\prod_{j=1}^{\ell} g_j(x)$	has all the α^i, $1 \leq i \leq \ell t$ as roots.
$(f(x))$	the principal ideal generated by f(x).
wt.$(f(x))$	the Hamming weight of f(x).
g(x)	the generator polynomial of a cyclic code.
n	the length of a cyclic code.
r	the number of check digits.

k	the number of information digits.
M	the period of a polynomial.
	the transmission rate.

A cyclic code is an ideal generated by g(x) in the residue class ring of polynomials mudulo x^M-1, where M is the least positive integer such that g(x) divides X^M-1. M is called the natural length of the cyclic code. If n=M, then the cyclic code is a nonshortened cyclic code. If n < M, then the cycli code is a shortened cyclic code.

(Definition 1) M is defined to be the period of f(x) if M is the least positive integer such that $f(x) \mid x^M-1$.

(Lemma 1) If N is the period of $\prod_{i=1}^{\ell} g_i(x)$, then the period of $\prod_{i=1}^{\ell} (g_i(x))^{2^{m_i}}$ is $2^m N$, where m=max. $(m_1, m_2, \ldots, m_\ell)$. (For the proof, the reader is referred to P. 151, Berlekamp (1968).)

(Lemma 2) The two polynomials $\frac{x^N+1}{g_1(x)g_2(x)}$ and $g_1(x)$ are relatively prime.

(Proof) If N is the order of α, then $N \mid 2^f-1$, and $X^N+1=0$ has N distinct roots, i.e., α^i, $1 \leq i \leq N$.

Suppose that

$$\frac{x^N+1}{g_1(x)\ g_2(x)}, \quad g_1(x) = m_1(x) \neq 1 \text{ and } \alpha^j \text{ is a root of } m_1(x),$$

then $x^N+1 = 0$ has at least a repeated root of α^j where $1 \leq j \leq N$ which is a contradiction. Therefore:

$$\frac{x^N+1}{g_1(x)\ g_2(x)}, \quad g_1(x) = 1. \qquad \text{Q.E.D.}$$

A direct result of lemma 2 is lemma 3. Lemma 3 is stated without proof.

(Lemma 3) The two polynomials $\frac{x^{2^m N}+1}{(g_1(x))^{2^m} g_2(x)}$ and $(g_1(x))^{2^m}$ are relatively prime.

(Lemma 4) g(x), where $g(x) = \prod_{j=1}^{\ell} g_j(x)$ generates a cyclic code of length N with minimum weight $\geq \ell t+1$. (For the proof, the reader if referred to Peterson (1961)).

(Lemma 5) If e(x) is the error polynomial of weight t and if $e(x) \equiv r(x)$ [mod. $x^\ell+1$], where deg. $(r(x)) < \ell$, then wt. $(r(x)) \leq$ wt. $(e(x))$.

III. CLASS I CYCLIC CODES

If $g_1(x)\ g_2(x)$ generates a basic BCH code of the length N with minimum

weight $\geq$ 2t+1, then $(g_1(x))^{2^m} g_2(x)$, where m is an arbitrary integer generates a class I cyclic code of the length $2^m N$ with minimum weight $\geq$ 2t+1.

(Theorem 1) Let g(x) be $(g_1(x))^2 g_2(x)$. g(x) generates a cyclic code of the length 2N with minimum weight $\geq$ 2t+1.

(Proof) Suppose that there exists a code polynomial a(x) $\neq$ 0 and wt. (a(x)) $\leq$ 2t. By using the Euclidean Algorithm, a(x) can be written as $a(x) = (x^N+1)$. Q (x)+r(x) where deg (r(x)) < N. As a result of lemma 5, it can be seen that wt. (r(x)) $\leq$ wt. (a(x)) $\leq$ 2t.

Since $g_1(x)\ g_2(x)$ | a(x) and $g_1(x)\ g_2(x)$ | x^N+1, $g_1(x)\ g_2(x)$ | r(x). r(x) is therefore a code polynomial of the length N. If r(x) $\neq$ 0, then wt. (r(x)) $\geq$ 2t+1, which is a contradiction. Therefore, $a(x)=(x^N+1)Q(x)$, where deg.Q(x) $\leq$ N-1. Since wt.(a(x)) $\leq$ 2t and deg. Q(x) < N, wt.(Q(x)) $\leq$ t.

Since $\dfrac{x^N+1}{g_1(x)g_2(x)}$, $g_1(x) = 1$ (lemma 2), $g_1(x)$ | Q (x).

Wt. (Q(x)) is therefore $\geq$ t+1, which is a contradiction. Therefore all the nonzero code vectors have a weight of at least 2t+1. Q.E.D.

(Theorem 2) $(g_1(x))^{2^m} g_2(x)$ generates a cyclic code of the length $2^m N$ with minimum weight $\geq$ 2t+1.

(Proof) Proof is by mathematical induction on the number m.

(1) For the m=0 case, $g_1(x)g_2(x)$ generates a BCH code of the length N with minimum weight $\geq$ 2t+1. For the m=1 case, theorem 1 shows that $(g_1(x)).g_2(x)$ generates a cyclic code of the length 2N with minimum weight $\geq$ 2t+1.

(2) Induction step

Suppose that $(g_1(x))^{2^m} g_2(x)$ generates a cyclic code of the length $2^m N$ with minimum weight $\geq$ 2t+1. Since $(g_1(x))^{2^m} g_2(x)$ | $(g_1(x)g_2(x))^{2^m}$, $(g_1(x)g_2(x))^{2^m}$ also generates a cyclic code of the length $2^m N$ with minimum weight $\geq$ 2t+1.

If $(g_1(x))^{2^{m+1}} g_2(x)$ generates a cyclic code of the length $2^{m+1} N$ with minimum weight $\leq$ 2t, then there exists a code polynomial, a(x) where a(x) $\neq$ 0 and wt. (a(x)) $\leq$ 2t.

By using the Euclidean Algorithm, $a(x) = (x^{2^m N}+1)$ Q (x)+r(x), where deg.r(x), deg Q(x) $\leq 2^m N$, wt. (r(x)) $\leq$ wt. (a(x)) $\leq$ 2t. Since $(g_1(x))^{2^m} g_2(x)$ | a(x) and $(g_1(x))^{2^m} g_2(x)$ | $x^{2^m N}+1$, $(g_1(x))^{2^m} g_2(x)$ | r(x). If r(x) $\neq$ 0, then by means of the induction hypothesis, wt. (r(x)) $\geq$ 2t+1, which is a contradiction. Therefore $a(x) = (x^{2^m N}+1)$ Q(x) where wt. (Q(x)) $\leq$ t.

Since $\dfrac{x^{2^m N}+1}{(g_1(x))^{2^m} g_2(x)}$, $(g_1(x))^{2^m} = 1$, $(g_1(x))^{2^m}$ | Q (x).

Consequently wt. (Q(x)) $\geq$ t+1, which is a contradiction.

Therefore all the nonzero code vectors must have a weight $\geq$ 2t+1. Q.E.D.

(Corollary 1) $g(x) = (g_1(x)\ g_2(x))^{2^m}$ generates a cyclic code of the length $2^m N$ with minimum weight $\geq$ 2t+1.

Throughout this paper, an equivalent BCH code is defined t be a BCH code which has the same lower bound of minimum weight and has roughly the same code length as the code in consideration. There is not a single binary Class I code which has a better efficiency than the equivalent BCH code.

If the exact minimum weight of the basic BCH code generated by $g_1(x) \cdot g_2(x)$ is 2t+1, then the exact minimum weight of the Class I Cyclic Code generated by $(g_1(x))^{2^m}\ g_2(x)$ is 2t+1. The result will be proved in theorem 3.

Theorem 3. If $g_1(x)\ g_2(x)$ generates a BCH code of the length N with minimum weight exactly equal to 2t+1, then $(g_1(x))^{2^m}\ g_2(x)$ generates a Class I Code of the length $2^m N$ with minimum weight exactly equal to 2t+1.

(Proof). It has been proved in Theorem 2 that the minimum weight of the Class I Code is at least 2t+1. It will be proved in the following that there exists a code polynomial of the Class I Code which is of weight 2t+1.

Since $g_1(x)\ g_2(x)$ generates a BCH code with minimum weight exactly equal to 2t+1, there exists a code polynomial a(x) which is of weight 2t+1. Since $g_1(x)\ g_2(x) \mid a(x)$, $g_1(x^{2^m}) \mid a(x^{2^m})$. Therefore $(g_1(x))^{2^m}\ g_2(x) \mid a(x^{2^m})$. Since $a(x^{2^m})$ is of degree $\leq 2^m N-1$, $a(x^{2^m})$ must be a code polynomial generated by $(g_1(x))^{2^m}\ g_2(x)$. Since wt. $(a(x^{2^m}))$ = wt. $(a(x))$ = 2t+1, there exists a code polynomial $a(x^{2^m})$ which is of weight 2t+1. Therefore, the minimum weight of the Class I Code is exactly equal to 2t+1.

Q.E.D.

IV. CLASS II SHORTENED CYCLIC CODES

The shortened cyclic code for multiple-error correction have been studied by Corr (1966). Class II codes are much more easily constructed and are a much wider class of codes than those reported in Corr (1966). In this section an algorithm for constructing the Class II shortened cyclic codes will be presented.

Algorithm for constructing Class II codes

If $h_i(x)$ generates a cyclic code of the length N_i with minimum weight $\geq 2^i$t+1, where $1 \leq i \leq \ell$, then $\prod_{i=1}^{\ell} h_i(x)$ generates a shortened cyclic code of the length $\sum_{i=1}^{\ell} N_i$ with minimum weight $\geq 2^{\ell}$ t+1 provided that g.c.d. $\dfrac{x^{N_i}+1}{h_i(x)}$, $\prod_{j=1}^{i} h_j(x) = 1$, and $N_i > \sum_{j=1}^{i-1} N_j$ for $2 \leq i \leq \ell$.

(Theorem 4) $g(x) = h_1(x)\ h_2(x)$ generates a shortened cyclic code of the length (N_1+N_2) with minimum weight $\geq 2^2$t+1.

(Proof) Suppose that there exists a non-zero code polynomial a(x) which has a weight $\leq 2^2$t. By using the Euclidean algorithm, $a(x) = (x^{N_2}+1)Q(x)$

+r(x), where deg (Q(x)) $\leq N_1-1 < N_2$, wt. (r(x)) $\leq 2^2t$ and deg (r(x)) $\leq N_2-1$. Since $h_2(x)$ divides both $x^{N_2}+1$ and a(x), $h_2(x)$ also divides r(x). r(x) is therefore a code polynomial. If $r(x) \neq 0$, then r(x) must have a weight $\geq 2^2t+1$, which is a contradiction. Therefore $a(x) = (x^{N_2}+1)\, Q(x)$, where wt. (Q(x)) $\leq 2t$. By assumption, g.c.d. $\dfrac{x^{N_2}+1}{h_2(x)}$, $h_1(x) = 1$, therefore $h_1(x) \mid Q(x)$. Q(x) is therefore a code polynomial and has a a weight $\geq 2^2t+1$. Q.E.D.

(Theorem 5) $\prod_{j=1}^{\ell} h_j(x)$ generates a shortened cyclic code of the length $\sum_{i=1}^{\ell} N_i$ with minimum weight $\geq 2^{\ell}t+1$.

(Proof) The proof of this theorem is by mathematical induction on the number ℓ.

(1) For the $\ell=1$ case, $h_1(x)$ generates a cyclic code of the length N_1 with minimum weight $\geq 2t+1$.

For the $\ell=2$ case, the previous theorem proves that $h_1(x)\, h_2(x)$ generates a shortened cyclic code of the length (N_1+N_2) with minimum weight $\geq 2^2t+1$.

(2) Induction step

It is assumed that $\prod_{j=1}^{\ell-1} h_j(x)$ generates a class II code of the length $\sum_{j=1}^{\ell-1} N_j$ with minimum weight $\geq 2^{\ell-1}t+1$.

Suppose that $\prod_{j=1}^{\ell} h_j(x)$ generates a shortened cyclic code of length $\sum_{i=1}^{\ell} N_i$ with minimum weight $\leq 2^{\ell}t$. Let a(x) be a non-zero code polynomial and wt. $(a(x)) \leq 2^{\ell}t$. By using the Euclidean Algorithm, $a(x) = (x^{N_\ell}+1)$ $Q(x) + r(x)$ where deg. $(Q(x)) \leq \sum_{i=1}^{\ell-1} N_i < N_\ell$ and wt. $(r(x)) \leq 2^{\ell}t$. If $r(x) \neq 0$, then $g(x) \mid r(x)$ and wt. $(r(x)) \geq 2^{\ell}t+1$, which is a contradiction. Therefore $r(x) = 0$ and $a(x) = (x^{N_\ell}+1)\, Q(x)) \leq 2^{(\ell-1)}t$.

Since $\dfrac{x^{N_\ell}+1}{h_\ell(x)}$, $\prod_{i=1}^{(\ell-1)} h_i(x) = 1$, $\prod_{i=1}^{\ell-1} h_i(x) \mid Q(x)$. Therefore wt. $(Q(x)) \geq 2^{\ell-1}t+1$, which is a contradiction.

Therefore every non-zero polynomial must have a weight $\geq 2^{\ell}t+1$. Q.E.D.

Two subclasses of codes are further defined for the Class II Codes.

(Definition 2) If $h_j(x)$ generates a Class I Code for the length $2^{i_j}N$ with minimum weight $\geq 2^jt+1$, where $\prod_{i=1}^{(j-1)} h_i(x) \mid h_j(x)$ and $i_\ell > i_{(\ell-1)} > \cdots > i_1 \geq 0$, $1 \leq i \leq \ell$, then the Class II Codes generated by $\prod_{j=1}^{\ell} h_j(x)$

are called the Class II-B Codes. The rest of the Class II Codes are called Class II-A Codes.

It can be easily seen that the conditions $N_i > \sum_{j=1}^{i-1} N_j$ and $\frac{N_{j+1}}{h_j(x)}$, $\prod_{i=1}^{j-1} h_i(x) = 1$ where $2 \leq i \leq \ell$ are satisfied by the Class II - B Codes.

Examples of Class II-A Codes and II-B Codes are given in Tables 1 and 3. Table 2 gives the equivalent BCH codes of the Class II-A Codes listed in Table 1. It can be seen from Table 1 and 2 that if $N_\ell \gg \sum_{i=1}^{(\ell-1)} N_i$, then the Class II-A Code has a rate of transmission comparable to the equivalent BCH code. However, there is no binary Class II-B Code which has an efficiency comparable to the equivalent BCH code.

V. CLASS III CYCLIC CODES

The Class III Cyclic Codes are obtained by modifying the Class II-B shortened cyclic codes. An algorithm for constructing the Class III Codes will be presented in the following:

Algorithm for Constructing Class III Codes.

If $h_1(x)$ generates a Class I Code of the length $2^i N$ with minimum weight $\geq 2t+1$ and $h_j(x)$ generates a class I code of the length $2^{(i+j-2)} N$ with minimum weight $\geq 2^j t+1$, where $\prod_{i=1}^{(j-1)} h_i(x) \mid h_j(x)$, $2 \leq j \leq \ell$, then $\prod_{j=1}^{\ell} h_j(x)$ generates a cyclic code of the length $2^{(i+\ell-1)} N$ with minimum weight $\geq 2^\ell t+1$.

If the exact minimum weight of the Class I code which is generated by $h_\ell(x)$ is $2^\ell t+1$, then the exact minimum weight of the Class III code generated by $\prod_{j=1}^{\ell} h_j(x)$ is $2^\ell t+1$. This result will be proved in Theorem 7. Table 4 gives a few examples of Class III Codes.

(Theorem 6) $\prod_{j=1}^{\ell} h_j(x)$ generates a cyclic code of the length $2^{(i+\ell-1)} N$ with minimum weight $\geq 2^\ell t+1$.

Since the proof of the theorem is similar to the proof of Theorem 4, the proof will not be supplied. It can be easily seen that the period $\prod_{j=1}^{\ell} h_j(x)$ is $2^{(i+\ell-1)} N$, therefore $\prod_{j=1}^{\ell} h_j(x)$ generates a cyclic code of the length $2^{(i+\ell-1)} N$ with minimum weight $2^\ell t+1$.

(Theorem 7) If $h_\ell(x)$ generates a Class I Code with minimum weight equal to $2^\ell t+1$, then $(\prod_{j=1}^{\ell} h_j(x))$ generates a class III cyclic code with

minimum weight exactly equal to $2^{\ell}t+1$.

(Proof) It has been proved in Theorem 6 that the minimum weight of the Class III cyclic code is at least $2^{\ell}t+1$. In the following, it will be shown that there exists a code polynomial which is exactly of weight $2^{\ell}t+1$.

Since $h_{\ell}(x)$ generates a cyclic code of the length $2^{i+\ell-2}N$ with minimum weight exactly $2^{\ell}t+1$, there exists a code polynomial $a(x)$ which is of weight $2^{\ell}t+1$. Since $\prod_{j=1}^{(\ell-1)} h_j(x) \mid h_{\ell}(x)$, $\prod_{j=1}^{\ell} h_j(x) \mid h_{\ell}(x^2)$. If $h_{\ell}(x) \mid a(x)$, then $h_{\ell}(x^2) \mid a(x^2)$. Therefore $\prod_{j=1}^{\ell} h_j(x) \mid a(x^2)$. Since $a(x^2)$ is a polynomial of degree $\leq (2^{i+\ell-1}N-2)$, $a(x^2)$ must be a code polynomial in the cyclic code generated by $\prod_{j=1}^{\ell} h_j(x)$. Since wt. $(a(x^2)) = 2^{\ell}t+1$, the minimum weight of the Class III code is exactly equal to $2^{\ell}t+1$.

Q.E.D.

ACKNOWLEDGMENT

The author is grateful to Dr. Berlekamp for the many helpful suggestions and in particular, for his excellent book (Berlekamp, 1968) from which the author obtained the knowledge to do this research.

REFERENCES

[1] Berlekamp, E.R., (1968), "Algebraic Coding Theory", McGraw-Hill Book Company

[2] Bose, R.C. and Ray-Chauhuri, D.K., (1960), "On a Class of Error Correcting Binary Group Codes," Inf. and Control, 3, 68-79.

[3] Corr, F., (1966), "A Shortened Cyclic Code for Multiple Errors", Trans. on Inf. Theory, IEEE vol. IT-12, No. 3.

[4] Peterson, W.W., (1961), "Error Correcting Codes", the M.I.T. Press, Cambridge, Mass.

Table 1 Examples of Class II-A Codes

N_1	$h_1(x)$	N_2	$h_2(x)$	N_3	$h_3(x)$	N_4	$h_4(x)$	n		d_{min}	g(x)
7	(13)*	15	(23)(37)	1023	(2011) (2017) X (2415) (3771)			1045	95%	9	$h_1(x)h_2(x)\ h_3(x)$
7	(13)(15)	15	(31)(23) X (37)	1023	(2011)(2017)(2415)X (3771)(2257)(2065) X(2157)(2653)			1045	90.6%	17	$h_1(x)h_2(x)h_3(x)$
15	(23)	31	(45)(75)	63	(103)(127)(147)X (111)	2047	(4005)(4445)X (4215)(4055)X (6215)(7413)X (4143)(4563)	2156	94.4%	17	$h_1(x)h_2(x)h_3(x)h_4(x)$

* The octal representation of the polynomial x^3+x+1.

Table 2 Equivalent BCH Codes

n	k	d_{min}	
1023	983	9	96%
1023	943	17	92%
2047	1059	17	95.4%
2047	1915	25	93.5%

Table 3 Examples of Class II-B Codes

i	The Binary Representation of i	n	$h_3(x)$	$h_2(x)$	$h_1(x)$	dmin	g(x)
3	(0011)	3N		$(g_1(x)g_2(x))^{2^X}$ $g_3(x)g_4(x)$	$g_1(x)g_2(x)$	4t+1	$h_1(x)h_2(x)$
5	(0101)	5N		$(g_1(x)g_2(x))^{2^{2^X}}$ $g_3(x)\ g_4(x)$	$g_1(x)g_2(x)$	4t+1	$h_1(x)h_2(x)$
6	(0110)	6N		$(g_1(x)g_2(x))^{2^{2^X}}$ $g_3(x)\ g_4(x)$	$(g_1(x))^2 g_2(x)$	4t+1	$h_1(x)h_2(x)$
7	(0111)	7N	$(g_1(x)g_2(x)g_3(x)g_4(x))^{2^{2^X}}$ $g_5(x)g_6(x)g_7(x)g_8(x)$	$(g_1(x)g_2(x))^{2^X}$ $g_3(x)g_4(x)$	$g_1(x)\ g_2(x)$	8t+1	$h_1(x)h_2(x)h_3(x)$

Where t is any arbitrary positive integer.

Table 4 Examples of Class III Codes

n	$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$	dmin	g(x)
4N	$g_1(x)g_2(x)$	$g_1(x)g_2(x)g_3(x)g_4(x)$	$(g_1(x)g_2(x)g_3(x)g_4(x))^2 X$ $g_5(x)g_6(x)g_7(x)g_8(x)$		8t+1	$h_1(x)h_2(x)h_3(x)$
8N	$(g_1(x))^2 g_2(x)$	$(g_1(x)g_2(x))^2 g_3(x)g_4(x)$	$(g_1(x)g_2(x)g_3(x)g_4(x))^{2^2} X$ $g_5(x)g_6(x)g_7(x)g_8(x)$		8t+1	$h_1(x)h_2(x)h_3(x)$
8N	$g_1(x)g_2(x)$	$g_1(x)g_2(x)g_3(x)g_4(x)$	$(g_1(x)g_2(x)g_3(x)g_4(x))^2 X$ $g_5(x)g_6(x)g_7(x)g_8(x)$	$(\prod_{j=1}^{8} g_j(x))^{2^2} X$ $(\prod_{j=9}^{16} g_j(x))$	16t+1	$h_1(x)h_2(x)h_3(x)h_4(x)$

MAXIMUM LIKELIHOOD AND SEQUENTIAL DECODING OF SHORT CONSTRAINT LENGTH CONVOLUTIONAL CODES

F. JELINEK*
L. R. BAHL
IBM, Thomas J. Watson Research Center
Yorktown Heights, N. Y.

ABSTRACT

A comparison is made between the probabilities of error and decoding speeds for the two decoding algorithms. The results of a search for good convolutional codes of short constraint length are also presented.

INTRODUCTION

As a result of the maximum likelihood decoding procedure suggested by Viterbi (1), there has recently been considerable interest in convolutional codes of short constraint length. The amount of storage and computation to do maximum likelihood decoding grows exponentially with constraint length and, therefore, this decoding procedure is practical for short constraint lengths only. Since maximum likelihood decoding is 'optimal', the performance of this decoding method should be superior to that of a sequential decoder.

This paper is a preliminary report of some results on the performance of short constraint length codes. More results will be published shortly in a report entitled 'An investigation of short constraint length convolutional codes'. The first part of the paper presents the results of a comparison between the two decoding methods on a binary symmetric channel for various codes of rate ½ and constraint lengths c=4,5,6 and 7. Probabilities of error and decoding speeds are compared. The second part of the paper presents the results of a search for good codes of short constraint length. Both systematic and non-systematic codes were tested. If $g_1=g_{11},g_{12},\ldots g_{1c}$ and $g_2=g_{21},g_{22},\ldots g_{2c}$ are the generators of a code, then the non-systematic codes tested all had the property that $g_{11}=g_{21}=g_{1c}=g_{2c}=1$.

The simulation was done on an IBM 360/91 computer. The programs were written in Fortran by ordinary, run-of-the-mill programmers,us, and no particular effort was made towards optimizing the programs. The maximum likelihood decoder was a straightforward mechanization of the Viterbi algorithm. A stack decoding (2) algorithm was implemented to perform sequential decoding. The frame length was 2050 bits, the number of information bits per frame being 1026-c.

One measure of performance of a code is its bit error rate p_b, i.e. the fraction of bits decoded wrong. However, the bit error rate does not completely reflect the intrinsic strength of a code. For any code, it is possible to change the

---

*On sabbatical leave from Cornell University, Ithaca, N.Y.

bit error rate by a reassignment of codewords. We define another measure of performance, the state error rate p_s which is more indicative of the strength of the code. For an introduction to the concept of state in convolutional codes, see Omura (3). A burst of b errors containing no run of c-1 or more zeros will cause the decoder to be in the wrong state for b+c-2 time intervals. The decoder output is, of course, unreliable for b time intervals, though the total number of bit errors may be less than b. The state error rate is the fraction of time the decoder output is unreliable. Both bit and state error rates are shown in the tables.

Other parameters of interest are the minimum distance d_m and the free distance d_f. It is believed that d_f influences the error rate, especially at low channel error probabilities, and d_m influences the depth of search into an incorrect subtree in sequential decoding (4).

COMPARISON OF THE TWO DECODING ALGORITHMS

Tables I-IV show the performance of some rate ½ codes of constraint length 4,5,6 and 7 at a channel probability of error .045 ($R_{comp}=\frac{1}{2}$). The bit and state error rates obtained with maximum likelihood decoding and sequential decoding are tabulated. The codes are listed in order of increasing bit error rate.

It is interesting to note that for any constraint length the code having the lowest bit error rate for maximum likelihood decoding also has the lowest bit error rate for sequential decoding. Furthermore, for c=4, there is no significant difference between the performance of the two decoders for any of the codes tested. As c increases, sequential decoder performance keeps up with maximum likelihood decoder performance for the best codes only. As the quality of the code (i.e. its performance for maximum likelihood decoding) deteriorates, maximum likelihood decoder performance becomes considerably superior to sequential decoder performance. For the best codes, the sequential decoder seems to do a very good job of finding the optimal path. The slight difference in error rates seems to come mainly from the random choice among several equally likely optimal paths. The state error rate exhibits more or less the same behavior as the bit error rate. It is important to note that the statements above are only for codes of constraint lengths 4,5,6 and 7. So far no comparison has been made for longer lengths.

The average decoding time per bit for sequential decoding at c=4 was .15 milliseconds which rises to .2 milliseconds for c=7. The corresponding times for maximum likelihood decoding were .16 milliseconds (c=4) and 1.1 milliseconds (c=7). The amount of time needed to do maximum likelihood decoding became too large to get statistics on codes of longer constraint length, though the performance of some longer codes was tested with sequential decoding. Since a general purpose computer is not particularly suitable for either of the two decoding methods, the decoding speeds achieved do not have any great significance.

COMPARISON OF CODES

At present there is no systematic method of constructing good convolutional codes. Some codes have been obtained by Bussgang (5) and Lin and Lyne (6) by a search procedure which maximizes the minimum distance. A large number of codes of constraint lengths 4,5,6,7,10 and 15 were tested at a channel error probability of .045. The results are summarized in Tables I-VI. Advantage was taken of certain symmetries in code generators, to reduce the number of codes to be tested. Interchanging g_1 and g_2 does not change the code properties in any way. Reversing both g_1 and g_2, i.e. taking the reciprocals of g_1 and g_2 as the generators, leaves code performance unchanged for maximum likelihood decoding, but may change sequential decoder performance. Codes which are not self-reciprocal were run both forwards and backwards for sequential decoding and the results of the better direction are shown in the tables. Codes exhibiting catastrophic failure (7) were rejected. One particularly interesting class of codes are the complement codes, where $g_{11}=g_{21}=g_{1c}=g_{2c}=1$ and g_{2i} is the complement of g_{1i} for $i\neq 1$ or c. The complement codes have large free distance which assures good code performance, at least at low channel error probabilities.

For c=4 and 5, all systematic codes and all codes having $g_{1c}=g_{2c}=1$ were tested. Beyond c=5, the number of codes grows too rapidly to allow for exhaustive testing. For c=6 and 7, all complement codes and other known good codes such as the Bussgang and Massey-Costello (4) codes were tested. For c=10 and 15, some of the known codes and some complement codes were tested. The complement codes were constructed by choosing as one of its generators an irreduceable polynomial of the desired length having approximately half zeros and half ones. Choosing an irreduceable polynomial assures us that the code is non-catastrophic.

Tests were also made at $p_c=.0225$ $(R/R_{comp}= .85)$ and $p_c=.09 (R/R_{comp}=1.45)$. Some of the results are shown in Table VII for c=5.

One interesting conclusion from the results obtained is that minimum distance is a very poor indicator of code performance. The free distance seems to be a better indicator, especially at low channel error probabilities. However, at higher channel error probabilities, it has less influence on bit and state error rates, and it is clear that no simple criteria can be used to predict the relative performance of codes.

State and bit error rates show more or less the same behavior. Non-systematic codes perform much better than systematic codes at and below R_{comp}. As constraint length increases, the difference between the best systematic and non-systematic codes found became greater. Above R_{comp}, the superiority of non-systematic codes vanishes (see Table VII). However, all the codes perform so poorly at $p_c=.09$ that their usefulness above R_{comp} is extremely doubtful. It is also seen from Table VII that the relative performance of codes changes with channel

error probability. Another interesting result was that at short constraint lengths, single and double errors do not predominate.

The complement codes seem to perform very well at all constraint lengths. For c=6 and 7, quite a few other codes (other than those shown in Tables III and IV) were tested, and none of them were found to perform better than the complement codes. One reason for their low error rate is, of course, the high value of free distance.

Further investigations into the structure of convolutional codes are being carried out. It is hoped that a theory will emerge to explain the behavior of these codes.

ACKNOWLEDGMENT

We wish to thank Prof. J. Massey of the University of Notre Dame for a very interesting and illuminating conversation. Some of his remarks on the significance of minimum and free distance are quoted in this paper. We also wish to thank him for supplying some codes and data from experiments at Notre Dame.

GLOSSARY OF SYMBOLS AND ABBREVIATIONS

c	constraint length
d_m	minimum distance
d_f	free distance
g_1, g_2	code generators
p_b	bit error rate
p_c	channel error probability
p_s	state error rate
M.L.D.	maximum likelihood decoding
S.D.	sequential decoding

REFERENCES

1. A. J. Viterbi, 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', IEEE Trans. Inf. Theory, IT-13, Apr. 1967.

2. F. Jelinek, 'A fast sequential decoding algorithm utilizing a stack', to appear in IBM Journal of Res. and Dev., vol. 13, Nov. 1969.

3. J. Omura, 'On the Viterbi decoding algorithm', IEEE Trans. Inf. Theory, IT-15, Jan. 1969.

4. J. L. Massey, private communication.

5. J. J. Bussgang, 'Some properties of binary convolutional code generators', IEEE Trans. Inf. Theory, IT-11, Jan. 1965.

6. S. Lin and H. Lyne, 'Some results on binary convolutional code generators', IEEE Trans, Inf. Theory, IT-13, Jan. 1967.

7. J. L. Massey and M. K. Sain, 'Inverses of linear sequential circuits' IEEE Trans. Computers, C-17, Apr. 1968.

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
1101 1111		4	6	.0045	.0043	.0058	.0057
1101 1001		3	5	.0047	.0047	.0080	.0081
1101 1011	Complement	3	6	.0067	.0086	.0064	.0082
1000 1101	Systematic Bussgang	4	4	.0121	.0111	.0154	.0141
1000 1001	Systematic	3	3	.0135	.0129	.0231	.0214
1000 1111	Systematic	3	4	.0139	.0132	.0166	.0155

Average decoding time/bit M.L.D.– .16 msec., S.D.– .15 msec.

Table I. Performance of codes for c=4, p_c=.045

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
10011 11011		4	7	.0030	.0031	.0053	.0046
10011 11101	Complement	3	7	.0031	.0033	.0042	.0047
10011 10111		3	7	.0032	.0032	.0056	.0051
10001 11111	Complement	4	6	.0033	.0033	.0045	.0047
10011 11001		4	6	.0034	.0032	.0054	.0045
10101 11101	Massey-Costello	3	6	.0038	.0037	.0073	.0063
11011 11111		4	6	.0040	.0039	.0058	.0050

Table II. (Continued on next page)

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
10111 11111		4	6	.0040	.0042	.0068	.0062
10011 10101		4	6	.0041	.0043	.0065	.0062
10001 10011		3	5	.0044	.0059	.0097	.0145
10101 11111		3	6	.0046	.0045	.0081	.0077
10011 11111		4	6	.0055	.0053	.0076	.0073
10001 10101		3	5	.0058	.0082	.0135	.0212
10000 11011	Systematic	4	5	.0065	.0121	.0100	.0155
10000 11101	Systematic Bussgang	4	5	.0067	.0122	.0104	.0171
10000 10011	Systematic	4	4	.0106	.0097	.0157	.0141
10000 10001	Systematic	3	3	.0136	.0143	.0295	.0281
10000 11111	Systematic	3	4	.0145	.0138	.0193	.0174
10000 10101	Systematic	3	4	.0175	.0159	.0294	.0262

Average decoding time/bit M.L.D.- .3 msec., S.D.- .15 msec.

Table II. Performance of codes for c=5, p_c=.045

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
101011 111011	Massey-Costello	4	7	.0024	.0028	.0036	.0043
110111 101001	Complement	4	8	.0025	.0028	.0037	.0043
110001 101111	Complement	4	8	.0028	.0031	.0043	.0049
100000 110101	Systematic Bussgang	5	5	.0048	.0068	.0076	.0107
100000 111011	Systematic Bussgang	5	6	.0070	.0065	.0099	.0093

Average decoding time/bit M.L.D.- .54 msec., S.D.- .18 msec.

Table III. Performance of codes for c=6, p_c=.045

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
1100111 1011001	Complement	5	9	.0012	.0013	.0017	.0020
1101111 1010001	Complement	5	9	.0012	.0013	.0020	.0020
1110111 1001001	Complement	5	8	.0013	.0014	.0018	.0021
1100101 1011011	Complement	4	9	.0013	.0015	.0018	.0022
1100001 1011111	Complement	5	8	.0014	.0017	.0019	.0027
1110001 1001111	Complement	5	8	.0014	.0015	.0021	.0025
1101011 1010101	Complement	5	8	.0015	.0015	.0022	.0022

Table IV. (Continued on next page)

g_1 g_2	Comment	d_m	d_f	M.L.D. p_b	S.D. p_b	M.L.D. p_s	S.D. p_s
1101001 1010111	Complement	4	8	.0017	.0020	.0025	.0030
1000001 1111111	Complement	4	6	.0030	.0041	.0053	.0070
1000000 1110011	Systematic Bussgang Adj.	5	6	.0049	.0049	.0075	.0078

Average decoding time/bit M.L.D.- 1.1 msec., S.D.- .2 msec.

Table IV. Performance of codes for c=7, p_c=.045

g_1 g_2	Comment	d_m	d_f	S.D. p_b	S.D. p_s
1001011001 1110100111	Complement	6	12	.00035	.00055
1100110001 1011001111	Complement	4	12	.00042	.00073
1010011001 1101100111	Complement	5	12	.00045	.00075
1010110111 1110110111	Massey-Costello	5	9	.00103	.00191
1000000000 1110011001	Systematic Bussgang	6	7	.00408	.00860
1000000000 1110011011	Systematic Bussgang Adj.	6	8	.00374	.00672

Table V. Performance of codes for c=10, p_c=.045

g_1 g_2	Comment	d_m	d_f	S.D. p_b	S.D. p_s
101001111110001 110110000001111	Complement	7	14	.000002	.000002
100000101100111 111111010011001	Complement	6	14	.000006	.000012
101011011101011 111011011101011	Massey- Costello	7	11	.00015	.00028
100000000000000 110111000010111	Systematic Bussgang	8	10	.00149	.00364
100000000000000 111001101100111	Systematic Bussgang	8	10	.00183	.00559
100000000000000 111101101000001	Systematic Bussgang	8	9	.00223	.00559

Table VI. Performance of codes for c=15, p_c=.045

g_1 g_2	p_c=.0225 p_b	p_c=.045 p_b	p_c=.09 p_b
10011 11101	.00014	.0031	.0562
10011 11011	.00020	.0030	.0500
10101 11101	.00026	.0038	.0494
10000 11011	.00092	.0065	.0458
10000 10111	.00097	.0067	.0459

Table VII. Performance of codes for c=5, p_c=.0225,.045,.09

QUASI-OPTIMAL CONTROL WITH NONPERIODIC PIECEWISE-CONSTANT FEEDBACK GAINS

LOUIS R. NARDIZZI
Information and Control Sciences Center
Institute of Technology
Southern Methodist University
Dallas, Texas 75222

ABSTRACT

The quasi-optimal control of nonlinear systems with nonquadratic performance indices over a finite-time interval is presented in this paper. The synthesis of nonperiodic piecewise-constant feedback gains is accomplished by gradient algorithms. A few examples are presented to demonstrate the validity of the proposed method.

INTRODUCTION

A great deal of attention has recently been directed to the synthesis of quasi-optimal (near optimal) feedback control systems [1,2,3]. A particular method proposed by Nahi and Bettwy [1] is applicable to nonlinear systems and nonquadratic performance indices. However, they synthesize a nonlinear feedback control law which may be difficult and expensive to implement for some applications. In certain engineering problems a linear feedback control law with nonperiodic piecewise-constant feedback gains will achieve a suitable measure of "closeness" to the desired performance of the original problem and still be reasonably simple to implement. The inherent difficulties in the original optimal feedback problem should be reduced in any quasi-optimal method and the dual task of reducing the computational requirements of the original problem and synthesizing an implementable control law is paramount. The method proposed in this paper is an attempt to accomplish this task.

The physical process to be controlled is defined by the vector differential equation

$$\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t) \quad \text{with} \quad \underline{x}(t_o) = \underline{x}_o \tag{1}$$

The state $\underline{x}(t)$ is an n-dimensional vector in E^n, the n-dimensional Euclidean space. $\underline{u}(t)$ is an r-dimensional control vector in the set Ω(open) and $\underline{f}$ is an n-dimensional vector which is in $C^{(1)}$ on $E^n \times \Omega$. The optimal open-loop control $\underline{u}^*(t)$ transfers the state of the system from $\underline{x}(t_o)$ to some desired final state $\underline{x}(t_f)$ in such a manner as to minimize the performance index

---

This research was supported in part by the U.S. Air Force Office of Scientific Research under Grant No. AF-AFOSR F44620-68-C-0023.

$$J = \int_{t_o}^{t_f} L(\underline{x}, \underline{u}, t)dt \tag{2}$$

Consider the difference between an optimal open-loop control and a closed-loop control defined as

$$\delta\underline{u}(t) \triangleq \underline{u}(x) - \underline{u}^*(t) \tag{3}$$

The optimal open-loop control and feedback control drive equation (1) from the same initial state. The difference in the final states due to the two controls can be approximated by a Taylor series expansion such that

$$\delta\underline{x}(t_f) = \int_{t_o}^{t_f} \phi(t,s)\left(\frac{\partial f}{\partial u}\right)\delta\underline{u}(s)ds \tag{4}$$

where $\phi(t,s)$ is the fundamental matrix solution of

$$\dot{\delta\underline{x}} = \left(\frac{\partial f}{\partial x}\right)\delta\underline{x} \tag{5}$$

with $\frac{\partial f_i}{\partial x_j}$, the i, jth element of the nxn matrix $\left(\frac{\partial f}{\partial x}\right)$ evaluated along the optimal open-loop trajectory. Taking the norm of both sides of equation (4) and applying Schwartz's Inequality yields

$$||\delta\underline{x}(t_f)||^2 \leq \left[\int_{t_o}^{t_f} ||\phi(t,s)(\frac{\partial f}{\partial u})||^2 ds\right]^{1/2}\left[\int_{t_o}^{t_f} ||\delta\underline{u}(s)||^2 ds\right]^{1/2} \tag{6}$$

The right-hand side of inequality 6 represents an upper bound on $||\delta\underline{x}(t_f)||^2$ and implies that a new performance index can be used to define a suitable measure of "closeness" to the optimal open-loop control. The only quantity on the right hand side of inequality 6 which depends on $\underline{u}(x)$ is $\delta\underline{u}(s)$ and, consequently, a new performance index I can be defined

$$I \triangleq \int_{t_o}^{t_f} ||\delta\underline{u}(s)||^2 ds \tag{7}$$

Assume, for simplicity, a scalar control u(x) which is linear in the states $\underline{x}(t)$ and that equation (1) represents a completely observable system. A linear control law is assumed in order to minimize computational requirements and achieve a reasonably simple implementable control. The control

u(x) is defined as

$$u(x) \triangleq \underline{\alpha}^T(t_k)\underline{x}(t, \underline{\alpha}(t_k)) \quad \text{for} \quad t_k \leq t < t_{k+1} \tag{8}$$

where $\alpha^T(t_k)$ represents the transpose of the feedback gain vector.

The ith component of the n-dimensional feedback gain vector $\underline{\alpha}(t_k)$ is defined by

$$\alpha_i(t_k) \triangleq m_i(t_k)[\eta(t-t_k) - \eta(t-t_{k+1})] \tag{9}$$

where $\eta(t)$ is the Heavyside step function and $m_i(t_k)$, the constant amplitude of the ith component of $\underline{\alpha}(t_k)$ during the time interval $t_k \leq t < t_{k+1}$. Equation 9 defines the nonperiodic piecewise-constant feedback gains. The new performance index can now be written in terms of the Euclidean norm as

$$I = \sum_{k=0}^{N} \int_{t_k}^{t_{k+1}} [\underline{\alpha}^T(t_k)\underline{x}(s, \underline{\alpha}(t_k)) - u^*(s)]^2 ds \tag{10}$$

The new performance index is minimized over the sequence of piecewise-constant feedback gains $\{\underline{m}(t_k), k = 0, 1, \ldots N\}$ and switching times $\{t_k, k = 1, 2, \ldots N\}$. The minimization of I over these parameters converts the original feedback problem to a minimum terminal miss problem. That is, the synthesis of n(N+1) + N parameters is required to solve the quasi-optimal control problem where N is a preselected number of switching times. N should remain small in order to keep implementation of the control law simple but can be enlarged to satisfy terminal miss requirements. If N = 0, there are no switching times but there are n amplitudes to synthesize. In the limit of N, the closed-loop and open-loop trajectories will be identical.

Necessary conditions can be derived for an extremal value of the new performance index I by setting the first variation of I with respect to amplitude feedback gains and switching times to zero [4]. These conditions are

$$\int_{t_k}^{t_{k+1}} [\underline{\alpha}^T(t_k)\underline{x}(s, \underline{\alpha}(t_k)) - u^*(s)][\underline{x}(s, \underline{\alpha}(t_k)) + \left(\frac{\partial x}{\partial \alpha}\right)^T \underline{\alpha}(t_k)]ds = \underline{0}$$

$$k = 0, 1, 2 \ldots N \tag{11}$$

and

$$\Delta_- u(t_k)[2u^*(t_k) - \Delta_+ u(t_k)] = 0 \qquad k = 1, 2, 3, \ldots N \tag{12}$$

$\left(\frac{\partial x}{\partial \alpha}\right)$ is the solution of the matrix differential equation (parameter sensitivity equation)

$$\left(\frac{\partial \dot{x}}{\partial \alpha}\right) = \left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial x}{\partial \alpha}\right) + \frac{\partial f}{\partial \alpha} \tag{13}$$

In addition,

$$\Delta_{-}u(t_k) = \underline{\alpha}^T(t_k)\underline{x}[t_k, \underline{\alpha}(t_k)] - \underline{\alpha}^T(t_{k-1})\underline{x}[t_k, \underline{\alpha}(t_{k-1})] \tag{14}$$

and

$$\Delta_{+}u(t_k) = \underline{\alpha}^T(t_k)\underline{x}[t_k, \underline{\alpha}(t_k)] + \underline{\alpha}^T(t_{k-1})x[t_k, \underline{\alpha}(t_{k-1})] \tag{15}$$

Equation 11 is a condition which will be used to synthesize the amplitude of the piecewise-constant feedback gains and equation 12 will be used to determine the switching times.

Synthesis Technique and Examples

A gradient algorithm based on equations 11 and 12 can be used to synthesize the quasi-optimal closed-loop controls for the problem presented. The step-by-step procedure follows for each fixed value of N.

1.) Compute the optimum open-loop control u*(t) by any of the standard methods (Maximum Principle, Gradient Techniques, Dynamic Programming, etc.). Store u*(t) and values of the states $\underline{x}^*(t_f)$ at the final time.

2.) Choose nominal values for the parameters $\{m^\rho(t_k),\ k = 0, 1, \ldots N\}$ and $\{t_k^\rho,\ k = 1, 2, \ldots N\}$ of the control $u^\rho(x)$ to derive equations 1 and 13 forward in time.

3.) Calculate the new values of the amplitude of the feedback gains and switching times according to the following gradient algorithms

$$\underline{m}^{\rho+1}(t_k) = \underline{m}^\rho(t_k) - c\int_{t_k^\rho}^{t_{k+1}^\rho} [\underline{\alpha}^T \cdot \underline{x} - u^*(s)][\underline{x} + \left(\frac{\partial x}{\partial \alpha}\right)^T \underline{\alpha}]ds \tag{16}$$

$$k = 0, 1, \ldots N$$

$$t_k^{\rho+1} = t_k^\rho - c\{\Delta_{-}u(t_k)[2u^*(t_k) - \Delta_{+}u(t_k)]\} \tag{17}$$

$$k = 1, 2, \ldots N$$

c is a positive constant appropriately chosen to speed the convergence [5].

4.) Use these parameters in the control $u^{\rho+1}(x)$ to drive equation 1 to a final state $\underline{x}^{\rho+1}(t_f)$ which satisfies the required terminal miss. (i.e.

$$||\underline{x}^{\rho+1}(t_f) - \underline{x}^*(t_f)|| \leq \delta_1$$

with δ_1 the suitably small terminal miss).

5.) If the terminal miss is not satisfied return to step 3 and continue. N should be selected small to begin with and increased if the terminal miss is unsatisfactory. A few examples have been attempted and the results thus far compare favorably with other near optimal methods [1,2].

The first example is that of a linear plant and nonquadratic cost function. The plant equations and initial conditions are

$$\begin{aligned} \dot{x}_1 &= x_2 & x_1(0) &= 2.0 \\ \dot{x}_2 &= -x_2 + u & x_2(0) &= 2.0 \end{aligned} \tag{18}$$

The cost function to be minimized is

$$J = \int_0^5 \left\{x_1^2 + x_2^4 + u^2\right\} dt \tag{19}$$

The minimum cost with open-loop control is 19.9191. The results for 0, 1, 2, and 3 switchings are shown in Table 1 for this example. Zero switchings require the synthesis of only two parameters. The two parameters are the amplitudes $m_1(0)$ and $m_2(0)$. The minimum cost for the closed-loop control is in close agreement with the optimum open-loop cost. This example shows that a linear, constant feedback gain control law is comparable to a control law with multiple switchings or a nonlinear feedback control law [1].

The second example is that of a nonlinear plant and quadratic cost. The plant dynamics are given by

$$\begin{aligned} \dot{x}_1 &= x_2 & x_1(0) &= 1.5 \\ \dot{x}_2 &= (1-x_1^2)x_2 - x_1 + u & x_2(0) &= 1.5 \end{aligned} \tag{20}$$

and the cost function is

$$J = \int_0^5 \left\{x_1^2 + x_2^2 + u^2\right\} dt \tag{21}$$

The minimum cost for the open-loop control is 8.78. Once again, the results for up to 3 switchings are shown in Table 2 and compare favorably with nonlinear control laws.

CONCLUSIONS

The approach taken in this paper is that of determining a quasi-optimal linear control law which gives a performance comparable to other quasi-optimal control laws but is simpler to implement. The problems of stability and sensitivity to initial conditions requires extensive investigation. The

examples attempted to date indicate that the method is feasible and should be studied further.

ACKNOWLEDGMENT

The author would like to thank Mr. Ming Tarng for valuable assistance in programming the results contained in this report.

REFERENCES

[1] Nahi, N. E. and Bettwy, T. S., "Synthesis of Near Optimal Feedback Control Systems," Hawaii International Conference on System Sciences, 29-31, January 1968.

[2] Kleinman, D. L., Fortmann, T. and Athans, M., "On the Design of Linear Systems with Piecewise-Constant Feedback Gains," IEEE Trans. on Automatic Control, Vol. AC-13, no. 4, August 1968, pp. 354-361.

[3] Rekasius, Z. V., "Suboptimal Design of Intentionally Nonlinear Controllers," IEEE Trans. on Automatic Control, Vol. AC-9, no. 4, October 1964, pp. 380-386.

[4] Nardizzi, L., "Suboptimal Control of a Class of Discrete-Continuous Systems," IEEE Trans. on Automatic Control, (to appear).

[5] Nardizzi, L., and Bekey, G. A., "Synthesizing Optimal Controls for Modulated Discrete-time Systems," Int. J. Control, 1968, Vol. 8, No. 6, pp. 571-590.

J	N	$m_1(0)$	$m_2(0)$	t_1	$m_1(t_1)$	$m_2(t_1)$	t_2	$m_1(t_2)$	$m_2(t_2)$	t_3	$m_1(t_3)$	$m_2(t_3)$
19.9499	0	-.9963	-1.1705									
19.9426	1	-.9803	-1.1156	2.1	-.3540	.0207						
19.9422	2	-.9803	-1.1156	1.93	-.3651	.0294	2.55	-.3541	.0207			
19.9422	3	-.9804	-1.1156	1.93	-.3652	.0294	2.53	-.3540	.0207	3.25	-.3540	.0207

Table 1 - Linear System and Nonquadratic Cost

J	N	$m_1(0)$	$m_2(0)$	t_1	$m_1(t_1)$	$m_2(t_1)$	t_2	$m_1(t_2)$	$m_2(t_2)$	t_3	$m_1(t_3)$	$m_2(t_3)$
9.4809	0	-.3996	-1.8027									
9.3895	1	-.3483	-1.6438	2.60	.3024	-2.3171						
9.1556	2	-.2604	-1.1830	2.00	.5716	-1.9545	4.25	-.7611	-.6725			
8.8106	3	-.3603	-.6610	.85	-1.2027	-2.8940	2.98	1.2349	-1.9474	4.13	-.3892	-.5110

Table 2 - Nonlinear System and Quadratic Cost

A CONCISE EXPRESSION FOR THE SWITCHING COEFFICIENTS FOR POSITIVE-NEGATIVE FEEDBACK SYSTEMS

JOHN PAUL HYDE
Frank J. Seiler Research Laboratory
USAF Academy, Colorado 80840

Abstract. The optimum switching condition for positive-negative feedback control systems is investigated. A compact algebraic formula for calculating the coefficients of the major switching equation for n-th order systems is derived, eliminating the cumbersome calculations required by previous methods. The formula is valid even when there are repeated negative-feedback poles.

I. INTRODUCTION

Several years ago Meiksin [1] introduced a unique feedback control system that operates in a regenerative positive-feedback mode during the first part of the transient response period and switches into negative feedback for the remaining part. Switching in this positive-negative (p-n) feedback control system is contrived to eliminate one component of the negative-feedback error transient, customarily the slowest of the dominant components. Both the period of positive feedback at the outset and the succeeding period of negative-feedback, with the slowest component removed, serve to speed-up the response. An added advantage suggested by Garrett [2] is that small inputs or disturbances result in only small signals that drive the system to zero. Thus, the p-n feedback system does not cause the severe chattering associated with "bang-bang" systems for example.

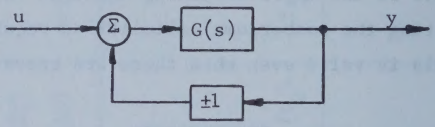
The optimum switching condition, developed by Meiksin [1] and Garrett [2] for second- and third-order p-n feedback systems is a linear combination of the system error and its time derivatives (another advantage over systems with non-linear switching conditions) with the weighting coefficients expressed in terms of the system's negative-feedback poles. However, these previous developments specifically excluded systems with repeated negative-feedback poles. Furthermore, the methods they indicated for determining coefficients of the switching expression for higher (n-th) order systems are computationally cumbersome, entailing solving for and then inverting an n-dimensional matrix and determining the system's n negative-feedback poles.

Section II of this paper reanalyzes the switching concept for n-th order p-n feedback systems in order to develop a concise algebraic expression for the coefficients of the optimum linear switching condition. The simple expression, derived in Section III, involves only constants from the system's negative-feedback characteristic polynomial and the value of the

one negative-feedback pole corresponding to the error transient component to be eliminated. When applied to second- and third-order systems, the new formula yields the results obtained by earlier authors. Finally, it is shown in Section IV that this generalized expression is valid even when there are repeated negative-feedback poles.

II. SWITCHING EXPRESSION FOR N-TH ORDER P-N FEEDBACK SYSTEMS

Consider the single-variable p-n feedback system shown below.



Simple p-n feedback system

We assume the time-invariant plant G(s) includes sufficient pure integrators to produce zero steady-state error in response to step-function inputs under negative feedback. That is, if the intended inputs are step-position functions we assume G(s) includes at least one pure integrator (Type 1 plant); if the intended inputs are step-velocity functions we assume G(s) includes at least two pure integrators (Type 2 plant); at least three pure integrators (Type 3 plant) for step-acceleration inputs; etc.

This system's transient response to step-function inputs is divided into two time intervals. During the first interval, which begins at $t = 0$, the system is in the positive feedback mode. The second interval begins at $t = t_s$ when the system switches into negative feedback. Only the negative-feedback mode must be analyzed to determine the switching expression.

Given the plant transfer function

$$G(s) = \frac{Kz(s)}{s^x p(s)}$$

where z(s) and p(s) are real polynomials and x is the number of pure integrators. If G(s) is a real physical system, the degree of z(s) will be less than n, the degree of $s^x p(s)$. In the negative-feedback mode, the plant is driven by the system error, $e = u - y$, and the error/input transfer function is

$$\frac{E(s)}{U(s)} = \frac{1}{1 + G(s)} = \frac{s^x p(s)}{f(s)} \qquad (1)$$

where

$$f(s)_- = s^x p(s) + Kz(s)$$
$$= s^n + b_{n-1}s^{n-1} + \ldots + b_1 s + b_0 \qquad (2)$$

is the negative-feedback characteristic polynomial.

From (1), $f(s)_- \cdot E(s) = s^x p(s) \cdot U(s)$. The inverse Laplace transform is

$$\overset{(n)}{e}(\tau) + b_{n-1}\overset{(n-1)}{e}(\tau) + \ldots + b_1\dot{e}(\tau) + b_0 e(\tau) = \mathcal{L}^{-1}[s^x p(s)\cdot U(s)] \qquad (3)$$

where $\tau = t - t_s$ and

$$\left.\begin{aligned} \dot{e} &= de/d\tau \\ &\vdots \\ \overset{(n)}{e} &= d^n e/d\tau^n \end{aligned}\right\} \qquad (4)$$

If the input u is some step-type function and x is sufficiently large, the right-hand side of (3) is identically zero; that is

$$\overset{(n)}{e}(\tau) + b_{n-1}\overset{(n-1)}{e}(\tau) + \ldots + b_2\ddot{e}(\tau) + b_1\dot{e}(\tau) + b_0 e(\tau) = 0 \qquad (5)$$

The relations (4) and (5) can be expressed in matrix form as

$$\dot{E} = B\cdot E \qquad (6)$$

where $E = \mathrm{col}[e\ \dot{e}\ \ddot{e}\ \text{---}\ \overset{(n-1)}{e}]$ and

$$B = \begin{bmatrix} 0 & | & & I & \\ \text{---} & | & \text{---} & \text{---} & \text{---} \\ -b_0 & | & -b_1 & -b_2 \ldots . & -b_{n-1} \end{bmatrix}$$

is the companion matrix of $f(s)_-$. Thus $f(s)_-$ is both the minimum polynomial and the characteristic polynomial of B, and its roots $\lambda_1 \ldots \lambda_n$, the negative-feedback poles, are the characteristic roots of B [3]. For the moment we will assume that the n roots λ_i are distinct and simple (non-repeated).

Equation (6) can be transformed into its normal coordinates by the linear transformation $E = M\cdot X$ where $X = \mathrm{col}[x_1 x_2 \text{---} x_n]$ is the error vector in normal coordinates, and M is the n-dimensional modal matrix which diagonalizes B; that is,

$$M^{-1}\cdot B\cdot M = \mathrm{diag}\,[\lambda_1 \lambda_2 \text{---} \lambda_n] = D \qquad (7)$$

Applying the transformation $E = M\cdot X$ in (6) and premultiplying by M^{-1} gives $\dot{X} = M^{-1}\cdot B\cdot M\cdot X = D\cdot X$. This set has n independent differential

equations of the form $\dot{x}_i = \lambda_i x_i$ with solutions $x_i(\tau) = x_i(0)\exp(\lambda_i \tau)$, i = 1, ...,n. These decay to zero exponentially with increasing time (providing Real $[\lambda_i] < 0$) and any $x_i(\tau)$ can be zero for all $\tau > 0$ only if $x_i(0) = 0$.

The system error $e(\tau)$ is the first element of E and thus, by the transformation E = M·X, it is a linear combination of the $x_i(\tau)$:

$$e(\tau) = M_{11}x_1(\tau) + M_{12}x_2(\tau) + \ldots + M_{1n}x_n(\tau)$$

The error can be reduced to zero quicker if that component $x_i(\tau)$ which decays to zero slowest can be eliminated altogether. The p-n feedback system does this by responding first in the positive-feedback mode and then switching into negative feedback when one of the x_i (i.e., the slowest one) is identically zero.

Each component x_i is expressed in terms of the error e and its time derivatives in the inverse transformation X = N·E, where $N = M^{-1}$ is also an n-dimensional matrix. With notation $N = [N_{ij}]$, the component x_i is given in the i-th row of X = N·E; that is

$$x_i = N_{i1}e + N_{i2}\dot{e} + \ldots + N_{in}e^{(n-1)} \tag{8}$$

If we wish to eliminate that component of the negative-feedback error associated with the pole λ_i, the system must switch into negative feedback when

$$0 \equiv x_i = N_{i1}e + N_{i2}\dot{e} + \ldots + N_{in}e^{(n-1)} \tag{9}$$

Thus, (9) is the linear equation of the switching surface. Garrett [2] has shown that the system under positive feedback will pass through the state necessary to satisfy (9). He has also shown that switching according to (9) is optimum in the sense that it results in minimum-time response for a p-n feedback system that employs one switching instance in the transient response period.

Note that (9) could be normalized with respect to N_{i1} and still be valid.

III. A GENERAL EXPRESSION FOR THE COEFFICIENTS OF THE SWITCHING EQUATION

To implement switching according to (9), the coefficients N_{ik}, k = 1, 2,....,n, which are the elements of the i-th row of the matrix N, must be obtained. Post-multiplying (7) by $N = M^{-1}$ yields N·B = D·N, or

$$\begin{bmatrix} N_{11}N_{12} & \cdots & N_{1n} \\ N_{21} & & \vdots \\ \vdots & & \\ N_{n1} & \cdots & N_{nn} \end{bmatrix} \cdot \begin{bmatrix} & 1 & & \\ & & \ddots & \\ 0 & & & 0 \\ \hline -b_0 & -b_1 & -b_2 \cdots & -b_{n-1} \end{bmatrix} =$$

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & 0 \\ 0 & & \lambda_n \end{bmatrix} \cdot \begin{bmatrix} N_{11}N_{12} & \cdots & N_{1n} \\ N_{21} & & \vdots \\ \vdots & & \\ N_{n1} & \cdots & N_{nn} \end{bmatrix} \qquad (10)$$

Each row of (10) is an equation of two n-dimensional vectors. The i-th row is:

$$[(-b_0N_{in})\ (N_{i1}-b_1N_{in})\ (N_{i2}-b_2N_{in}) \cdots (N_{i,n-1}-b_{n-1}N_{in})] =$$

$$[(\lambda_iN_{i1})\ (\lambda_iN_{i2})\ (\lambda_iN_{i3}) \cdots (\lambda_iN_{in})] \qquad (11)$$

Equating corresponding elements of these two vectors:

$$-b_0N_{in} = \lambda_iN_{i1} \qquad (12)$$

$$N_{ij}-b_jN_{in} = \lambda_iN_{i,j+1} \qquad j = 1,2,\ldots,n-1 \qquad (13)$$

Without loss of generality, it is possible to set $N_{i1} = 1$. This essentially normalizes both the matrix equation (10) and the switching equation (9) with respect to N_{i1}. Then using $-N_{in}/\lambda_i = 1/b_0$ from (12) in (13) yields

$$N_{i,j+1} = \frac{N_{ij}-b_jN_{in}}{\lambda_i} = \frac{b_j}{b_0} + \frac{N_{ij}}{\lambda_i} \qquad j = 1,2,\ldots,n-1$$

Thus, the switching equation coefficients are completely defined by

$$\left.\begin{aligned} N_{i1} &= 1 \\ N_{i,j+1} &= \frac{b_j}{b_0} + \frac{N_{ij}}{\lambda_i} \qquad j = 1,2,\ldots,n-1 \end{aligned}\right\} \qquad (14)$$

This expression involves only constants b_i from the system's negative-feedback characteristic polynomial, readily determined from the plant transfer function as in (2), and the value λ_i of the one negative-feedback pole corresponding to the error transient component to be eliminated. This constitutes an improvement over earlier expressions

which required knowing the values of all the negative-feedback poles.

We can use those earlier expressions, however, to check the validity of (14). For example, Garrett [2] obtains the switching equation which will eliminate the component x_1 of a third-order system whose negative-feedback poles are known to be $s = \lambda_1$, λ_2, and λ_3. Then the system's negative-feedback characteristic polynomial is

$$f(s)_- = s^3 - (\lambda_1 + \lambda_2 + \lambda_3)s^2 + (\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1)s - \lambda_1\lambda_2\lambda_3$$

Here $b_2 = -(\lambda_1 + \lambda_2 + \lambda_3)$, $b_1 = (\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1)$ and $b_0 = -\lambda_1\lambda_2\lambda_3$. Using these in (14), the switching equation (9) is

$$0 \equiv x_1 = e - \frac{\lambda_2+\lambda_3}{\lambda_2\lambda_2}\dot{e} + \frac{1}{\lambda_2\lambda_3}\ddot{e}$$

which duplicates Garrett's result (his equation 30).

Morgan [4] has derived alternate though somewhat longer expressions for the elements of the characteristic row, but normalized with respect to the last element in the row (his equation 68). When normalized with respect to his formula for the first element, his expressions can be reduced to (14).

IV. THE CASE OF REPEATED NEGATIVE-FEEDBACK POLES

It will now be shown that the general expression (14) for the switching equation coefficients is valid even when there are repeated negative-feedback poles. Only the case where there is one double negative-feedback pole will be considered because the further extension of this proof to the more complex cases with triple, quadruple, etc., poles or with more than one repeated pole is straightforward.

If there is one double negative-feedback pole, the matrix B has the same double characteristic root. Then there exists a non-singular matrix $\underline{M}$ which reduces B to the Jordan canonical form

$$J = \underline{M}^{-1}\cdot B\cdot\underline{M} = \begin{bmatrix} \lambda_1 & & & & & 0 \\ & \ddots & & & & \\ & & \lambda_{k-2} & & & \\ & & & \lambda_k & 1 & \\ & & & & \lambda_k & \\ & & & & & \lambda_{k+1} \\ 0 & & & & & & \ddots \\ & & & & & & & \lambda_n \end{bmatrix}$$

where λ_k is the repeated pole (i.e., $\lambda_{k-1} = \lambda_k$) [5].

Applying the transformation $E = \underline{M}\cdot X$ in (6) and then premultiplying

by $\underline{M}^{-1}$ gives $\dot{X} = \underline{M}^{-1} \cdot B \cdot M \cdot X = J \cdot X$ or

$$\begin{bmatrix} \dot{x}_1 \\ | \\ \dot{x}_{k-2} \\ \dot{x}_{k-1} \\ \dot{x}_k \\ \dot{x}_{k+1} \\ | \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & 0 & \\ & & \lambda_{k-2} & & & & \\ & & & \lambda_k & 1 & & \\ & & & & \lambda_k & & \\ & & & & & \lambda_{k+1} & \\ & 0 & & & & & \ddots \\ & & & & & & \lambda_n \end{bmatrix} \begin{bmatrix} x_1 \\ | \\ x_{k-2} \\ x_{k-1} \\ x_k \\ x_{k+1} \\ | \\ x_n \end{bmatrix}$$

This set has n-1 independent differential equations of the form $\dot{x}_i = \lambda_i x_i$ with solutions $x_i(\tau) = x_i(0)\exp(\lambda_i \tau)$, $i = 1,2,\ldots,k-2,k,\ldots,n$ and one differential equation $\dot{x}_{k-1} = \lambda_k x_{k-1} + x_k$ with solution

$$\begin{aligned} x_{k-1}(\tau) &= [x_{k-1}(0) + \tau \cdot x_k(0)] \exp(\lambda_k \tau) \\ &= x_{k-1}(0) \exp(\lambda_k \tau) + \tau \cdot x_k(\tau) \end{aligned} \tag{16}$$

Note that two error transient components, x_{k-1} and x_k, have the same exponential-decay time constant, $1/\lambda_k$, the reciprocal of the repeated pole.

As before, we wish to switch from positive feedback into negative feedback to completely eliminate the error transient component associated with one negative-feedback pole. Switching to eliminate a component x_i corresponding to one of the non-repeated poles λ_i $(i \neq k-1,k)$ is accomplished by switching when $x_i \equiv 0$ $(i \neq k-1,k)$. However, switching to eliminate one of the two components associated with the repeated pole λ_k can only be accomplished by switching when $x_k \equiv 0$. By so doing, the component x_k will be zero for all subsequent time and the other component, x_{k-1}, associated with the repeated pole remains as a simple exponential:

$$x_{k-1}(\tau) = x_{k-1}(0)\exp(\lambda_k \tau) + \tau \cdot x_k(\tau) = x_{k-1}(0)\exp(\lambda_k \tau)$$

On the other hand, switching when $x_{k-1} \equiv 0$ does not completely eliminate any component: since $\tau \triangleq 0$ when the system switches into negative feedback, switching when $x_{k-1} \equiv 0$ causes $x_{k-1}(0) = 0$ but for $\tau > 0$, from (16), $x_{k-1}(\tau) = \tau \cdot x_k(\tau) \neq 0$.

Consequently, the system can eliminate any one negative-feedback error transient component x_i, except for $i = k-1$, by switching when $x_i \equiv 0$. Each of these components can be expressed in terms of the system error and its derivatives by the inverse transformation $X = N \cdot E$ where $N = \underline{M}^{-1}$.

In this transformation the i-th row is the same as (8). Postmultiplying (11) by $\underline{M}^{-1} = N$ yields

$$N \cdot B = \begin{bmatrix} \lambda_1 & & & & & & \\ & \searrow & & & & & \\ & & \lambda_{k-2} & & & & \\ & & & \lambda_k & 1 & & 0 \\ & & & & \lambda_k & & \\ & & & & & \lambda_{k-1} & \\ & 0 & & & & \searrow & \\ & & & & & & \lambda_n \end{bmatrix} \cdot N$$

As in (10) each row of this equation is an equation of two n-dimensional vectors. The i-th row (for $i \neq k-1$) is identical to (11) from which the general expression (14) for the switching coefficients N_{ij} was obtained. Consequently, (14) is valid even in this case.

Incidentally, there is no possibility of inadvertently calculating the coefficients $N_{k-1,j}$ of the (k-1)-th row of N since the general expression for those coefficients would be different from expression (14) because of the off-diagonal element 1 in the (k-1)-th row of the Jordan canonical matrix (15).

V. SUMMARY AND CONCLUSIONS

The optimum switching condition for p-n feedback control systems has been investigated. The major contribution of this effort is a compact algebraic formula for calculating the coefficients of the switching equation for n-th order systems. The formula is shown to be valid even when there are repeated negative-feedback poles.

Switching in these systems is contrived to eliminate one component of the negative-feedback error transient, customarily the slowest of the dominant components, by switching from positive into negative feedback when equation (9) is satisfied. The coefficients N_{ij} of the switching variable x_i in (9) are given by (14) in which the b_j are coefficients of the negative-feedback characteristic polynomial, equation (2), and λ_i is the negative-feedback pole corresponding to the transient component to be eliminated (i.e., λ_i is the algebraically largest of the dominant negative-feedback poles).

This simple formula for the switching coefficients should significantly reduce the need for tedious, exacting computations during the synthesis or analysis of p-n feedback systems, especially in the early exploratory phases. For example, a designer might quickly obtain approximate values

of the dominant negative-feedback poles using a method such as given by Chen [6]. With these approximate values he could select the negative-feedback error transient component to be eliminated, and then determine the exact value of that one pole by any one of several common techniques, such as iterated synthetic division. With that value and the negative-feedback characteristic polynomial, all coefficients of the switching variable can be determined easily, thereby significantly shortening design or analysis time.

REFERENCES

1. Meiksin, Z.H., "Positive-Feedback Phase-Space Trajectories and Application to Servo Systems," AIEE Transactions, Pt II, Vol 77, No. 40, pg 673-679, Jan 59.

2. Garrett, S.J., "A Linear Switching Condition for Third-Order Positive-Negative Feedback Control Systems," AIEE Transactions, Pt II, Vol 80, No. 54, pg 53-59, May 61.

3. Perlis, S. Theory of Matrices. Reading, Massachusetts: Addison-Wesley Publishing Co., Inc., 1958.

4. Morgan, B.S., Jr., "Sensitivity Analysis and Synthesis of Multivariable Systems," IEEE Transactions on Automatic Control, Vol AC-11, No. 3, pg 506-512, Jul 66.

5. Bellman, R. Introduction to Matrix Analysis. New York: McGraw-Hill Book Co., 1960.

6. Chen, K., "A Quick Method for Estimating Closed-Loop Poles of Control Systems," AIEE Transactions, Part II, Vol. 76, No. 30, pg. 80-87, May 1957.

APPLICATION OF GALERKIN'S METHOD FOR THE DETERMINATION OF TIME-OPTIMAL CONTROL OF DISTRIBUTED-PARAMETER SYSTEMS*

S. S. PRABHU and I. McCAUSLAND
Department of Electrical Engineering, University of Toronto, Toronto 181, Canada

ABSTRACT

Attention has been drawn to the Method of Weighted Residuals (MWR) for modelling distributed-parameter systems (DPS) described by partial differential equations with mixed initial and boundary conditions. Galerkin's method, a particular form of the MWR, has been considered for modelling DPS described by scalar partial differential equations. An illustrative time-optimal control example involving a one-dimensional heat conduction equation with a single boundary input has been solved utilizing Galerkin's method. Comparison with published work shows that the method yields a good approximation to the optimal control for the problem considered.

I INTRODUCTION

A powerful tool for determining approximate solutions to large classes of partial differential and integro-differential equations is the method of weighted residuals (MWR).[1] + The MWR comprises the following five basic methods: (1) method of collocation, (2) method of subdomains, (3) Galerkin's method, (4) method of moments, and (5) least-squares method. An attractive feature of the MWR is that it is applicable to linear and non-linear initial, boundary and eigenvalue problems. In this paper, application of Galerkin's method for the determination of approximations to time-optimal control of certain distributed-parameter systems (DPS) will be studied.

In general, there are two approaches for numerical solution of optimal control problems for DPS, both involving approximation: (1) the optimal problem is formulated for the equations describing the DPS, conditions for optimality are obtained and an approximate solution to the optimal control policy is obtained utilizing the conditions for optimality, (2) the DPS is approximated by a lumped-parameter system (LPS) by some technique, the specified performance index is approximated to one specified for the LPS model and an optimal control policy determined for the LPS. In this approach it is necessary to choose a LPS model which will yield an optimal control law for the LPS which is sufficiently close to the optimal control law for the DPS. The MWR for DPS falls into category (2).

Butkovskii,[2] who was among the first to consider solution by method (2), suggested using differential-difference

---

* This work was supported by the Canadian Commonwealth Scholarship and Fellowship Plan and by the National Research Council of Canada.

+ The numbers in square brackets designate references at the end of the paper.

equation approximations to DPS. This approach usually results in a high order LPS model for the DPS with its attendent numerical difficulty, especially if the problem involves 2 or 3-dimensional space geometry. McCausland [3], Wang and Tung[4] and Kitamori[5] obtain a LPS model using the truncated eigenfunction expansion of the DPS solution. This approach is limited to problems in which time and space variables can be separated. Even in situations where separation of time and space variables is possible, the eigenvalue problem obtained may be difficult to solve analytically. The MWR can be used for these situations. Moreover the MWR has been successfully applied in the past to obtain approximate solutions to non-linear problems. It is felt that a systematic study to exploit the potentialities of the MWR for the determination of approximations to optimal control policies for DPS is required.

II GALERKIN'S METHOD FOR INITIAL VALUE PROBLEMS

Consider a DPS defined by the non-homogeneous, scalar partial differential equation

$$N(\Theta) = f(\underline{x},t) \tag{1}$$

where $\Theta(\underline{x},t)$ is the state variable of the system, $\underline{x} = (x_1,x_2,x_3)$ defines a point in the 3-dimensional spatial domain Ω, $t \in [0,\infty)$ denotes the time coordinate, N is a partial differential operator, linear or non-linear, and $f(\underline{x},t)$ is the distributed input to the system. The non-homogeneous boundary conditions are of the form

$$L_i(\Theta) = u_i(\underline{x},t), \quad \underline{x} \in \Omega_b, \quad (i = 1,2,..,r) \tag{2}$$

where L_i, $(i = 1,2,..,r)$ are partial differential operators, linear or non-linear, defined on the boundary Ω_b of the spatial domain, and $u_i(\underline{x},t)$, $(i = 1,2,..,r)$ are the boundary inputs to the DPS. Initial conditions for the DPS are

$$\left.\frac{\partial^i \Theta}{\partial t^i}\right|_{t=0} = \Theta_o^i(\underline{x}), \quad (i = 0,1,2,..,n-1) \tag{3}$$

where $\frac{\partial^0 \Theta}{\partial t^0} \triangleq \Theta(\underline{x},t)$ and n is the order of the highest order time derivative of $\Theta(\underline{x},t)$ appearing in (1). Assuming the problem to be well-posed, consider an approximate representation of the solution of (1) under conditions (2) and (3) in the form

$$\Theta(\underline{x},t) \simeq \Theta^*(\underline{x},t) = \sum_{i=1}^{r} q_i(u_i(\underline{x},t),\underline{x},t) + \sum_{k=1}^{m} a_k(t)\phi_k(\underline{x}) \tag{4}$$

In (4), $\phi_k(\underline{x})$, $(k = 1,2,..,m)$ are linearly independent functions called coordinate functions and $a_k(t)$, $(k = 1,2,..,m)$ are unknown coefficients to be determined. The functions $\phi_k(\underline{x})$, $(k = 1,2,..,m)$ and $q_i(u_i(\underline{x},t),\underline{x},t)$, $(i = 1,2,..,r)$ are such that $\Theta^*(\underline{x},t)$ satisfies the boundary conditions defined by (2) exactly. The nature of the approximating function $\Theta^*(\underline{x},t)$ and the scheme to determine $a_k(t)$, $(k = 1,2,..,m)$ should be such that as m is increased, successively better approximations are obtained, and in the limit, as $m \to \infty$, $\Theta^*(\underline{x},t) \to \Theta(\underline{x},t)$, the exact solution, in some sense. Now,

the differential equation residual $R(\theta^*)$ and initial value residuals $R_o^k(\theta^*)$, $(k = 0,1,2,..,n-1)$ are formed as follows:

$$R(\theta^*) = N(\theta^*) - f(\underline{x},t) \tag{5}$$

$$R_o^k(\theta^*) = \theta_o^k(\underline{x}) - \frac{\partial^k}{\partial t^k}(\theta^*(\underline{x},t))\Big|_{t=0} \quad (k = 0,1,2,..,n-1) \tag{6}$$

where $\dfrac{\partial^o}{\partial t^o}\theta^*(\underline{x},t)\Big|_{t=0} \triangleq \theta^*(\underline{x},t)\Big|_{t=0}$

Galerkin's method to determine $a_k(t)$, $(k = 1,2,..,m)$ consists of making the residuals $R(\theta^*)$ and $R_o^j(\theta^*)$, $(j = 0,1,..,n-1)$ orthogonal to $\phi_k(\underline{x})$, $(k = 1,2,..,m)$ over Ω, that is,

$$\langle \phi_\ell, R(\theta^*)\rangle = 0, \quad (\ell = 1,2,..,m) \tag{7}$$

$$\langle \phi_\ell, R_o^j(\theta^*)\rangle = 0, \quad \left.\begin{matrix}(\ell = 1,2,..,m)\\ (j = 0,1,..,n-1)\end{matrix}\right\} \tag{8}$$

where $\langle \omega, v\rangle = \int_\Omega \omega v \, d\Omega$ defines the inner product notation used in (7) and (8). The set of equations (7) yields m ordinary differential equations for the m unknowns $a_k(t)$, $(k = 1,2,..,m)$. The m x n initial conditions necessary to solve these equations are given by (8). Thus the solution of (7) and (8) for $a_k(t)$, $(k = 1,2,..,m)$ and substitution in (4) will yield an approximate solution to the problem considered. The set of equations (7) with initial conditions given by (8) can be converted into m x n first order ordinary differential equations, which can be considered as an approximate state-equation obtained by Galerkin's method for the DPS under consideration.

III GALERKIN'S MODEL FOR THE PROBLEM CONSIDERED

For purposes of illustration, consider the one-dimensional heat-conduction equation

$$\frac{\partial\theta(x,t)}{\partial t} - \frac{\partial^2\theta(x,t)}{\partial x^2} = 0, \quad \begin{matrix}x \in (0,1),\\ t \in [0,\infty)\end{matrix} \tag{9}$$

with initial condition

$$\theta(x,t)\Big|_{t=0} = \theta_o(x) = 1.0 \tag{10}$$

and boundary conditions

$$\theta(x,t)\Big|_{x=0} = u(t); \quad \frac{\partial\theta(x,t)}{\partial x}\Big|_{x=1} = 0 \tag{11}$$

The boundary control input u(t) is a temperature which is magnitude constrained as follows:

$$|u(t)| \leqslant 1 \tag{12}$$

An approximate solution of the form

$$\theta^*(x,t) = \sum_{k=1}^{m} a_k(t)\, x^{k-1} \tag{13}$$

is assumed[+]. At any time instant, say t_1, the right hand side of (13) is a power series in x and from Weierstrass's theorem[7] it is known that $\exists$ a positive integer m such that the right hand side of (13) can approximate the true solution $\theta(x,t_1)$ arbitrarily closely in the sense of L_1 norm provided $\theta(x,t_1)$ is a continuous function of x. Thus (13) appears to be a reasonable form of the approximating function. Now, making $\theta^*(x,t)$ of (13) satisfy the boundary conditions given by (11), the following form for $\theta^*(x,t)$ is obtained:

$$\theta^*(x,t) = u(t) + \sum_{k=3}^{m} a_k(t)\left[x^{k-1} - (k-1)x\right] \qquad (14)$$

Redefining a_k's and the order of approximation m, and using the notation

$$\phi_k(x) = \left[x^{k+1} - (k+1)x\right] \qquad (15)$$

(14) can be written as

$$\theta^*(x,t) = u(t) + \sum_{k=1}^{m} a_k(t)\,\phi_k(x) \qquad (16)$$

Equation (16) is of the form of equation (4). Orthogonalizing the coordinate functions and again redefining a_k's, (16) can be written as

$$\theta^*(x,t) = u(t) + \sum_{k=1}^{m} a_k(t)\,\psi_k(x) \qquad (17)$$

where
$$\langle \psi_i, \psi_j \rangle = \begin{cases} 0 & i \neq j \\ \neq 0 & i = j \end{cases} \qquad (18)$$

Now evaluating the differential equation residual $R(\theta^*)$ as indicated in (5), and making the residual orthogonal to the coordinate functions $\psi_k(x)$, $(k = 1,2,..,m)$ as indicated in (7), and using vector-matrix notation, the following equation is obtained:

$$A\dot{\underline{a}} = B\underline{a} + C\frac{du}{dt} \qquad (19)$$

where
$$\begin{aligned} A &= \left[\langle \psi_i, \psi_j \rangle\right], \text{ an m x m matrix,} \\ B &= \left[\langle \psi_i, \frac{d^2\psi_j}{dx^2} \rangle\right], \text{ an m x m matrix,} \\ C &= \mathrm{Col}\,[c_1, c_2, .., c_m], \quad c_j = -\langle \psi_j, 1 \rangle \\ \underline{a} &= \mathrm{Col}\,[a_1, a_2, .., a_m] \end{aligned} \qquad (20)$$

From (19) the "state-equation" of the system

$$\dot{\underline{a}} = A^{-1}B\underline{a} + A^{-1}C\frac{du}{dt} \qquad (21)$$

can be written.

Determination of the initial condition for $\underline{a}$

Evaluating the initial condition residual $R_o(\theta^*)$ and making it orthogonal to the coordinate functions $\psi_k(x)$,

---

+ This form of approximation was considered for the first time by Bickley, [6].

(k = 1,2,..,m), as indicated in (6) and (8) the initial condition for $\underline{a}(t)$ is obtained to be

$$\underline{a}(0) = -A^{-1}C\left[1 - u(0)\right] \tag{22}$$

IV DETERMINATION OF TIME-OPTIMAL CONTROL

The time-optimal control problem for the DPS under consideration is to determine u(t) satisfying constraint condition (12), which will transfer the system from its initial state given by (10) to the final state

$$\Theta(x,t)\Big|_{t = t_f} = 0 \tag{23}$$

in minimum time t_f. Using the transformation

$$\underline{y} = \underline{a} - A^{-1}Cu \tag{24}$$

the "state-equation" (21) can be transformed to

$$\dot{\underline{y}} = A^{-1}B\underline{y} + A^{-1}BA^{-1}Cu \tag{25}$$

Using (22), the initial condition for $\underline{y}$ is $\underline{y}(0) = -A^{-1}C$ (26)

The final desired values for $\underline{a}(t)$ and u(t) are given by (17) and (23) to be

$$\underline{a}(t)\Big|_{t = t_f} = \underline{0} \; ; \; u(t)\Big|_{t = t_f} = 0 \tag{27}$$

Thus from (24), the final condition for $\underline{y}(t)$ is

$$\underline{y}(t)\Big|_{t=t_f} = \underline{0} \tag{28}$$

The time-optimal control problem for (25) which is an approximation to the optimal problem for the DPS under consideration is to determine u(t) subject to constraint (12) such that $\underline{y}(t)$ is transferred from its initial condition given by (26) to the origin in minimum time t_f. Now, the eigenvalues of $A^{-1}B$ are given as the solution of

$$\det\left[B - \lambda A\right] = 0 \tag{29}$$

A is a symmetric, positive-definite matrix; B is a symmetric matrix. It can be shown that (see [8] , pp 75-76) under these conditions the eigenvalues of $A^{-1}B$ are real. From Pontryagin's maximum principle it therefore follows that time-optimal control consists of switching u(t) between +1 and -1 at most (m-1) times and a last switching which brings u(t) to zero.

Determination of Switching Instants

Let the m switching instants for u(t) be $T_1, T_2, .., T_m$. Assuming u(0) = -1, the Laplace transform of u(t) is

$$\mathcal{L}\left[u(t)\right] = \bar{u}(s) = \frac{1}{s}\left[-1 + 2\varepsilon^{-T_1 s} - 2\varepsilon^{-T_2 s} + \right.$$
$$\left. + \ldots\ldots + 2(-)^m \varepsilon^{-T_{m-1} s} + (-)^{m+1}\varepsilon^{-T_m s}\right] \tag{30}$$

Let T be an m x m non-singular matrix such that

$$T^{-1}(A^{-1}B)T = \Lambda = \text{diag}\left[\lambda_1, \lambda_2, .., \lambda_m\right] \tag{31}$$

where $\lambda_1, \lambda_2, .., \lambda_m$ are eigenvalues of $A^{-1}B$. Transforming the state variables $\underline{y}$ to $\underline{y}^*$ according to

$$\underline{y} = T\underline{y}^* , \tag{32}$$

(25) can be transformed to

$$\dot{\underline{y}}^* = \Lambda\underline{y}^* + \Lambda H u , \tag{33}$$

where $$H = T^{-1}A^{-1}C = \text{Col}\left[h_1, h_2, .., h_m\right] \tag{34}$$

Also, $$\underline{y}^*(0) = T^{-1}\underline{y}(0) = -H \tag{35}$$

and $$\underline{y}^*(t_f) = \underline{0} \tag{36}$$

Taking the Laplace transform of (33) and simplifying using (35),

$$\bar{\underline{y}}^*(s) = (sI - \Lambda)^{-1}\Lambda H\bar{u}(s) - (sI - \Lambda)^{-1}H \tag{37}$$

Simplifying (37),

$$\bar{\underline{y}}^*(s) = \begin{bmatrix} \frac{h_1}{(s-\lambda_1)}\left[\lambda_1\bar{u}(s) - 1\right] \\ \frac{h_2}{(s-\lambda_2)}\left[\lambda_2\bar{u}(s) - 1\right] \\ \vdots \\ \frac{h_m}{(s-\lambda_m)}\left[\lambda_m\bar{u}(s) - 1\right] \end{bmatrix} \tag{38}$$

For the determination of time-optimal switching instants, the technique used by Goldwyn et al[9] is used. Since $\underline{y}^*(t) = \underline{0}$ for $t \geq T_m$, $\bar{\underline{y}}^*(s)$ cannot have singularities in the finite complex s-plane. (See [9], Appendix). It follows then from (38) that the equations

$$\left[\lambda_j\bar{u}(\lambda_j) - 1\right] = 0 , \quad (j = 1,2,..,m) \tag{39}$$

should be satisfied. Substituting for $\bar{u}(s)$ from (30), (39) can be written as

$$\left[-2 + 2\varepsilon^{-T_1\lambda_j} - 2\varepsilon^{-T_2\lambda_j} + \ldots + (-)^m 2\varepsilon^{-T_{m-1}\lambda_j} + (-)^{m+1}\varepsilon^{-T_m\lambda_j}\right] = 0$$

$$(j = 1,2,..,m) \tag{40}$$

(40) constitute m simultaneous algebraic equations for the m unknown switching instants and can be solved for them. In a problem if it is not certain whether u(0) = +1 or -1, equations similar to (40) have to be established for both u(0) = +1 and -1, both sets are to be solved and the set which yields minimum time control chosen.

V COMPUTATIONAL RESULTS

Table 1 gives eigenvalues computed for Galerkin's state-equations for orders of approximation m = 1,2,..,6 and also the first six eigenvalues obtained by the separation of variables technique. It is observed that the "dominant" eigenvalues of Galerkin's state-equations approximate very closely the corresponding ones obtained by the separation of variables technique and that the higher the value of m, the better the approximation. Table 2 gives the time-optimal switching instants for m = 1,2,..,6. The problem considered in this paper is the same as the first one considered in [9]. It is observed that the switching instants agree very closely with those of [9]. (See Table 1, [9]). The agreement is better, the higher the value of m.

VI REMARKS ABOUT THE NUMERICAL WORK INVOLVED

1. For the application of Galerkin's method, it is not necessary to orthogonalize the coordinate functions. But for the problem considered, if the functions are not orthogonalized, the inner-product matrix A defined for these functions is ill-conditioned, especially for higher values of m, and inversion of it to sufficient accuracy becomes difficult.

2. The functions $\phi_k(x)$, (k = 1,2,..,m) were orthogonalized to obtain $\psi_k(x)$, (k = 1,2,..,m) by Gram-Schmidt orthogonalization procedure[10] and in double-precision arithmetic.

3. Eigenvalues were determined using EIG5 Fortran subroutine for non-symmetric real matrices[11].

4. The simultaneous equations for the switching times were solved by generalized Newton-Raphson method. The starting point for an m th order approximation was obtained from the results for (m-1) th order approximation.

VII CONCLUSIONS

In this paper attention has been drawn to the MWR for obtaining approximate solutions to optimal control problems for DPS. Galerkin's method, a particular form of MWR has been considered in detail for modelling DPS described by scalar partial differential equations. The procedure can be easily extended to vector partial differential equations. Galerkin's method is shown to be capable of yielding lumped-parameter "state-equations" to DPS to which the known results of optimal control theory can be applied. For illustration, the method is applied to determine approximations to the time-optimal control of a DPS involving a one-dimensional heat conduction equation with a single boundary input. The results agree very well with those given in [9] in which the same problem has been solved by another method. Power series expansion has been used to derive coordinate functions. This is not the only possibility. One could choose trigonometric functions or Chebyshev polynomials or any other set of linearly independent functions, provided they form a complete set and the approximate solution $\theta^*(\underline{x},t)$ in terms of these functions satisfies the boundary conditions of the problem.

Approximations to time-optimal control for the boundary-controlled one-dimensional heat conduction equation with other types of linear boundary conditions have been solved by the authors (for example, for the case where the control input is heat flux instead of surface temperature, as in problem 2 of [9]) and the results are very satisfactory.

REFERENCES

1. B. A. Finlayson and L. E. Scriven, "The method of weighted residuals--a review," Applied Mechanics Reviews, vol. 19, pp.735-748, September 1966.
2. A. G. Butkovskii, "Some approximate methods for solving problems of optimal control of distributed parameters," Automation and Remote Control, vol. 22, No. 12, pp.1429-1438, June, 1962.
3. I. McCausland, "On optimum control of temperature distribution in a solid," Journal of Electronics and Control, vol. 14, pp.655-668, June, 1963.
4. P. K. C. Wang and F. Tung, "Optimal control of distributed parameter systems," Proc. of JACC, University of Michigan, pp.16-32, 1963.
5. Toshiyuki Kitamori, "Transformation of distributed-parameter systems into lumped-parameter systems for studies of optimum control," Proc. of the Third IFAC, London, pp.24A.1-24A.9, June, 1966.
6. W. G. Bickley, "Experiments in approximating to the solution of a partial differential equation," Philosophical Magazine, vol. 32, pp.50-56, 1941.
7. "Methods of numerical approximation", edited by D. C. Handscomb, pp.16-20, Pergamon Press, 1966.
8. Stephen H. Crandall, "Engineering analysis," McGraw-Hill Book Co. Inc., 1956, New York.
9. R. M. Goldwyn, K. P. Sriram and M. Graham, "Time-optimal control of a linear diffusion process", Journal SIAM Control, vol.5, No.2, pp.295-308, 1967.
10. Herbert Meschkowski, "Series expansions for mathematical phycisists", pp.61-66, Oliver and Boyd, 1968.
11. EIG5 subroutine for the determination of the eigenvalues of real matrices--"SHARE" library.

m	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
1	-2.50					
2	-2.4680441	-23.562725				
3	-2.4674060	-22.401209	-69.631383			
4	-2.4674011	-22.215410	-63.737864	-148.37934		
5	-2.4674011	-22.207052	-61.880514	-129.87753	-271.71968	
6	-2.4674011	-22.206621	-61.708097	-122.22432	-226.31002	-455.24749
Separation of variables	-2.4674010	-22.206609	-61.685026	-120.90265	-199.85948	-298.55552

Table 1

m	T_1	T_2	T_3	T_4	$-T_5$	T_6
1	0.27725887					
2	0.31253875	0.34192879				
3	0.31695251	0.35944272	0.36863202			
4	0.31809691	0.36375543	0.37901206	0.38294463		
5	0.31849676	0.36538253	0.38275682	0.38993960	0.39193222	
6	0.31865666	0.36609748	0.38450697	0.39310670	0.39701307	0.39813811

Table 2

OPTIMIZATION OF NONLINEAR CHARACTERISTICS

M. K. INAN and C. A. DESOER
Department of Electrical Engineering and Computer Sciences
Electronics Research Laboratory
University of California, Berkeley 94720

Abstract

The problem of minimizing a performance index of a finite dimensional dynamical system with respect to a single-input single-output, memoryless, and time-invariant nonlinearity $N(\cdot)$ within the system is considered. If $N(\cdot)$ belongs to a family of continuous functions with derivatives between + S and - S then the necessary conditions for the optimal nonlinearity $N^*(\cdot)$ roughly say that its derivative be + S or - S.

Introduction

The problem of optimizing a performance index with respect to a nonlinear characteristic has been a neglected area of research. A special case of this problem in which the system considered does not involve dynamics is discussed by Willems [4].

In this paper we state necessary conditions of optimality in a very general framework covering finite-dimensional dynamical systems. The proofs of the results will be found in [3].

Formulation of the Problem

We are given the following differential equation

$$\dot{x}(t) = f(x(t),N(cx(t))) \tag{1}$$

$x(t)$ = col. n-vector for each $t \in [0,T]$, $T > 0$ fixed.

$x(0) = x_0$ fixed.

c = constant row n-vector.

$f(\cdot,\cdot) \in \text{Lip } C^{(1)}$.

Definition $N(\cdot) \in \mathcal{N}$ iff

1]. $N(\cdot)$: $[a,b] \to R$, $N(0) = 0$, $a < 0$ and $b > 0$ real constants.

2]. $|N(y_1)-N(y_2)| \leq S|y_1-y_2|$, $S > 0$ a constant, $\forall y_1,y_2 \in [a,b]$.

3]. $\frac{dN(y)}{dy}$ is a piecewise continuous function† defined everywhere, except possibly a finite subset, on (a,b).

We define derivative from right and derivative from left of a real

†A function mapping R into R is said to be piecewise continuous iff on each finite interval it has finite number of discontinuities and it is regulated ([1] page 139).

valued function $N(\cdot)$ on R as follows,

$$\frac{d^{+}N(y)}{dy} \triangleq \lim_{\substack{\gamma>0 \\ \gamma\to 0}} \frac{N(y+\gamma)-N(y)}{\gamma} \tag{2a}$$

$$\frac{d^{-}N(y)}{dy} \triangleq \lim_{\substack{\gamma>0 \\ \gamma\to 0}} \frac{N(y)-N(y-\gamma)}{\gamma} \tag{2b}$$

Then, for $N \in \mathcal{N}$, it can be shown that $\frac{d^{+}N(y)}{dy}$ and $\frac{d^{-}N(y)}{dy}$ exist for all y in [a,b) and (a,b] respectively.

Remark The interval [a,b] is chosen such that all solutions of (1) for $N \in \mathcal{N}$ have the property

$$a < cx(t) < b \tag{3}$$

This is possible by the uniform Lipschitz constant S (also a bound for absolute value of the derivative of $N(\cdot)$) for all $N \in \mathcal{N}$.

By putting L^{∞} norm on $\mathcal{N}$ it becomes a subset of the Banach Space $C[a,b]$[†] [2]. This enables us to consider local minima.

We now define a functional P as follows

$$P(N) \triangleq \int_0^T h(x_N(t))dt \tag{4}$$

where $h(\cdot)\colon R^n \to R$, $h(\cdot) \in C^{(1)}$. $x_N(t)$ denotes the solution of (1) corresponding to $N \in \mathcal{N}$.

The basic optimization problem can now be stated as

$$\min_{N\in\mathcal{N}} P(N) \tag{5}$$

Statement of Theorems

Theorem 1 (Main Theorem) Suppose N* furnishes a local minimum for P for the basic problem given by (5). Furthermore assume that

---

[†] The set of all real valued continuous functions on [a,b] with the sup norm.

$$\mathcal{T} \triangleq \{t; \dot{cx}^*(t) = 0,\ t \in [0,T]\} = \{t_j\}_{j=1}^{m} \tag{6}$$

is a finite set. Under these conditions for all z (except possibly a finite set) in [a,b] the following relations hold.

$$\text{For } z \geq 0, \quad \frac{d^+N^*(z)}{dz} = -\,S\,\operatorname{sgn}\left[\int_{I^+(z)} \lambda^{T*}(\tau)\left[\frac{\partial f}{\partial N}\right]_{opt} d\tau\right], \tag{7a}$$

$$\text{For } z \leq 0, \quad \frac{d^-N^*(z)}{dz} = +\,S\,\operatorname{sgn}\left[\int_{I^-(z)} \lambda^{T*}(\tau)\left[\frac{\partial f}{\partial N}\right]_{opt} d\tau\right], \tag{7b}$$

where

$$I^+(z) = \{\tau; cx^*(\tau) > z\} \tag{8a}$$

$$I^-(z) = \{\tau; cx^*(\tau) < z\} \tag{8b}$$

and $\lambda^*(\tau)$ solves the adjoint equation

$$\dot{\lambda}^*(t) = -\left[\frac{df}{dx}\right]^T_{opt} \lambda^*(t) - \left[\frac{dh}{dx}\right]^T_{opt} \tag{9}$$

with $\lambda^*(T) = 0$.

Theorem 2 (Application) Consider the feedback system given by the equation

$$\dot{x} = Ax - bN(cx) \tag{10}$$

Assume that the linear time-invariant system (A, b, c) is completely observable.

Under these conditions if $x^*(t)$ denotes the optimum trajectory corresponding to N^* then either the assumption given by (6) in Theorem 1

† $\operatorname{sgn} \alpha \triangleq \begin{cases} 1 \text{ if } \alpha > 0 \\ -1 \text{ if } \alpha < 0 \\ \text{undefined if } \alpha = 0 \end{cases}$

A NEAR-OPTIMAL CONTROLLER FOR NONLINEAR SYSTEMS

MOHAMMAD JAMSHIDI
University of Illinois
Urbana, Illinois

ABSTRACT

A near-optimal controller for nonlinear systems is introduced which approximates the performance index up to (2n+1)th-order for an nth-order truncation of Taylor's series of control as in the optimally adaptive control. Nth-order sensitivity equations are derived from Maximum principle as in the optimally sensitive control. Two illustrative examples are included.

I. INTRODUCTION

The general purpose of an optimal control problem is to find a control ,u(t), which minimizes a given cost functional. Due to the difficulties encountered in optimization of nonlinear systems by the classical methods, i.e. gradient, Newton's Second Variation, etc., near-optimal controller designs have received particular attention in recent years. To name a few of these near-optimal techniques, "equivalent linearization," by Pearson [1], "parameter optimization," by Durbeck [2], "perturbation," by Al'Brekht [3], approximate solution to the Hamilton-Jacobi-Bellman's equation by Garrard, et. al. [4]. Another type of control laws called "optimally adaptive," by Werner and Cruz [5] minimizes the performance index regardless of the value(s) of system parameters or initial conditions. The use of mathematical sensitivity functions in a neighborhood of parameters and initial conditions has also had an important role in developing near-optimal control laws [6,7]. The "optimally sensitive" controller introduced by Kokotović, et. al. [7] is a first order approximation of optimally adaptive which presents a convenient derivation of sensitivity equations from Maximum Principle.

This paper will extend the techniques of optimally sensitive control to the nonlinear systems and preserves the same order of approximation of the performance as in the optimally adaptive.

Consider any nonlinear system whose state equations can be written as

$$\dot{\underline{x}} = \underline{g}(\underline{x},u) + \varepsilon \underline{f}(\underline{x},\underline{u}) \tag{1a}$$

$$\underline{g}(\underline{x},\underline{u}) = A\underline{x} + B\underline{u} \tag{1b}$$

where $\underline{x}$ is s-dimensional state, $\underline{u}$ is r-dimensional control, $\underline{g}$ and $\underline{f}$ are s-dimensional linear and nonlinear functions of $\underline{x}$ and $\underline{u}$, respectively. A is an sxs matrix, B is an sxr matrix which are functions of time in general. Parameter, $\varepsilon \varepsilon [\varepsilon_o, \varepsilon_f]$ introduces nonlinearities in the linear system $\dot{x} = g(\underline{x},\underline{u})$. For $\varepsilon = \varepsilon_o = 0$, the system is linear and called "nominal" and for $\varepsilon = \varepsilon_f$, it is termed "original." The control problem is to design a controller which produces a suboptimal control and eliminates difficulties being encountered in optimization of the "original" system. Furthermore, this suboptimal control approximately minimizes a performance index,

$$J = L(\underline{x},(t_f)) + \int_{t_o}^{t_f} V(\underline{x},\underline{u},t)\,dt \tag{2}$$

where L and V are scalar functions of quadratic form, t_o and t_f are fixed and $x(t_f)$ is free. The improvements in the performance and response characteristics of a closed loop controller over an open loop is demonstrated by two examples. Figure 1 shows the block diagram of the nonlinear plant with an undesigned controller.

II. NEAR-OPTIMAL CONTROLLER

System equation (1) is rewritten as

$$\dot{\underline{x}}(t,\varepsilon) = \underline{h}(\underline{x},\underline{u},t) = A\underline{x}(t,\varepsilon) + B\underline{u}(t,\varepsilon) + \varepsilon\underline{f}(\underline{x},\underline{u}) \tag{3a}$$

$$\underline{x}(t_o,\varepsilon) = \underline{x}^o \,. \tag{3b}$$

The cost functional (2) is written more explicitly as

$$J = \tfrac{1}{2}(\underline{x}'(t_f)M\underline{x}(t_f)) + \tfrac{1}{2}\int_{t_o}^{t_f}(\underline{x}'Q_1\underline{x} + \underline{u}'R_1\underline{u})\,dt \tag{4}$$

where M, Q_1, and R_1 are diagonal positive definite matrices and prime denotes transposition. The Hamiltonian function for the system (3) and performance (2) is of the form

$$H = -V + \underline{p}'\underline{h} \tag{5}$$

where p is an s-dimensional costate vector. The necessary conditions for optimality given by the Maximum Principle are

$$\dot{\underline{x}} = \nabla_p H \tag{6a}$$

$$\dot{\underline{p}} = -\nabla_x H \tag{7a}$$

$$0 = \nabla_u H \tag{8}$$

with boundary conditions,

$$\underline{x}(t_o,\varepsilon) = \underline{x}^o \tag{6b}$$

$$\underline{p}(t_f) = -\nabla_x L\big|_{t=t_f} . \tag{7b}$$

Let us introduce the following sensitivity vectors of the control, state, and costate,

$$\underline{\nu}_k = \left.\frac{\partial^k \underline{u}}{\partial \varepsilon^k}\right|_{\varepsilon=\varepsilon_o}, \quad \underline{\xi}_k = \left.\frac{\partial^k \underline{x}}{\partial \varepsilon^k}\right|_{\varepsilon=\varepsilon_o}, \quad \underline{\eta}_k = \left.\frac{\partial^k \underline{p}}{\partial \varepsilon^k}\right|_{\varepsilon=\varepsilon_o}, \quad k=1,2,3,\ldots,n.$$

for the sake of terminology, ν_k, for example is called "kth variation" control sensitivity function evaluated at $\varepsilon=\varepsilon_o$, similarly for ξ_k and η_k. It is assumed that r = 1, (one-dimensional control), g, f, V, and H are at least twice differentiable with respect to x,u, and x,p, and u all have continuous partial derivatives at $\varepsilon = 0$. Differentiating (6), (7), and (8) with respect to ε (assuming that all partial derivatives exist) and setting $\varepsilon_o = 0$ the first variation sensitivity functions equations are

derived

$$\dot{\underline{\xi}}_1 = h_x\,\underline{\xi}_1 + \underline{h}_u\,\underline{\nu}_1 + \underline{f} \tag{9}$$

$$\dot{\underline{\eta}}_1 = -H_{xx}\,\underline{\xi}_1 - \underline{h}_x\,\underline{\eta}_1 - H'_{xu}\,\underline{\nu}_1 - \underline{f}'_x\,\underline{P} \tag{10}$$

$$0 = H_{ux}\,\underline{\xi}_1 + h_u\,\underline{\eta}_1 + H_{uu}\,\underline{\nu}_1 + \underline{f}'_u\,\underline{P}\ . \tag{11}$$

After the elimination of $\underline{\nu}_1$ in (9), (10), and (11), the following is obtained

$$\dot{\underline{\xi}}_1 = F\,\underline{\xi}_1 + S\underline{\eta}_1 + \underline{\omega}_1 \tag{12a}$$

$$\dot{\underline{\eta}}_1 = Q\,\underline{\xi}_1 - F'\underline{\eta}_1 + \underline{\pi}_1 \tag{13a}$$

where,

$$Q = H_{ux}\,H_{uu}^{-1}\,H'_{ux} - H_{xx}\Big|_{\varepsilon=0} = Q_1$$

$$F = h_x - h_u\,H_{uu}^{-1}\,H'_{ux}\Big|_{\varepsilon=0} = A$$

$$S = -h_u\,H_{uu}^{-1}\,h'_u\Big|_{\varepsilon=0} = -BR_1^{-1}B'$$

$$\underline{\omega}_1 = \underline{f} - h_u\,H_{uu}^{-1}\,f_u\underline{P}\Big|_{\varepsilon=0} = \underline{\omega}_1^{\,o}$$

$$\underline{\pi}_1 = H'_{xu}\,H_{uu}^{-1}\,f'_u\,P - f_x P\Big|_{\varepsilon=0} = \pi_1^{\,o}$$ where H_{uu} is pos. def.

with boundary conditions,

$$\underline{\xi}_1(t_o) = \left.\frac{\partial \underline{x}}{\partial \varepsilon}\right|_{t=t_o} = \underline{0} \tag{12b}$$

$$\underline{\eta}(t_f) = \left.\frac{\partial \underline{p}}{\partial \varepsilon}\right|_{t=t_o} = \underline{0}\ . \tag{13b}$$

It is known [8,9] that $\underline{\eta}_1$ and $\underline{\xi}_1$ are related by the following relation,

$$\underline{\eta}_1 = -K\,\underline{\xi}_1 + \underline{g}_1 \tag{14}$$

where K is the solution to matrix Riccati equation [10],

$$\dot{K} = -KF - F'K + KSK - Q \tag{15}$$

with $K(t_f) = \nabla_{xx}\,L(x(t_f))$ and g_1 is the solution of the first variation adjoint system, evaluated at $\varepsilon = 0$,

$$\dot{\underline{g}}_1 = (KS - F')\underline{g}_1 + K\,\underline{\omega}_1 + \underline{\pi}_1 \tag{16a}$$

whose boundary condition in view of (12b), (13b), and (14) is,

$$\underline{g}_1(t_f) = K(t_f)\,\underline{\xi}_1(t_f)\,. \tag{16b}$$

Note that both equations (15) and (16) have to be solved backward in time since only their final conditions are known.

Substitution of $\underline{\eta}_1$ of (14) in (12) results,

$$\dot{\underline{\xi}}_1 = (F - SK)\underline{\xi}_1 + \underline{\alpha}_1 \tag{17}$$

where

$$\underline{\alpha}_1 = S\underline{g}_1 + \underline{\omega}_1 \ .$$

Considering sensitivity function $\underline{\nu}_1$ in (11) and (14), the following is evident,

$$\underline{\nu}_1 = -H_{uu}^{-1}(H'_{ux} - h'_u K)\underline{\xi}_1 + \underline{\beta}_1 \tag{18a}$$

where

$$\underline{\beta}_1 = -H_{uu}^{-1}(h_u g_1 + f'_u P) = T\underline{g}_1 + \underline{\theta}_1 \ . \tag{18b}$$

Note that (17) and (18) are the first variation sensitivity equations. The second variation equations can be similarly derived by differentiating (17) with respect to $\mathcal{E}$. The homogeneous portion of the new sensitivity equations remain the same and only the forcing terms, α_1 and β_1 have to be evaluated. The sub-optimal control and state consist of a truncated Taylor's series in parameter $\mathcal{E}$.

1. Open-Loop Design. Truncated Taylor Series expansion for control is termed open-loop, $u_{o\ell}$,

$$\underline{u}_{o\ell} = \underline{u}^* + \delta\underline{u}_{o\ell} \tag{19a}$$

$$\delta u_{o\ell} \cong \sum_{k=1}^{n} \frac{\mathcal{E}^k}{k!} \nu_k \ . \tag{19b}$$

Where $\underline{u}^*$ is the nominal control (i.e. $\mathcal{E}=\mathcal{E}_o=0$) and $\delta u_{o\ell}$ is the approximated difference between the open loop and nominal controls. For $\mathcal{E}_o = 0$, (1) will be reduced to the "nominal" system,

$$\dot{\underline{x}}^* = A\underline{x}^* + B\underline{u}^* . \tag{20}$$

The exact optimal control for (20) is the well-known linear state-regulator form (assuming the states are realizable), [10] which has the form,

$$\underline{u}^* = -R_1^{-1}B'K\underline{x}^* \ . \tag{21}$$

Where R_1, defined in (4) is matrix $H_{uu}|_{\mathcal{E}=0}$, and K is the solution of (15) at $\mathcal{E}=0$. Figure 2a shows a block diagram for the nominal system. The open loop controller (19) is substituted in the nonlinear system (3) and the result is

$$\dot{\underline{x}}_{o\ell} = A\underline{x}_{o\ell} + B(\underline{u}^* + \delta\underline{u}_{o\ell}) + \mathcal{E}\underline{f}(\underline{x}_{o\ell}, \underline{u}^* + \delta\underline{u}_{o\ell}) . \tag{22}$$

Figure 2(b) shows the open-loop controller diagram for nth variation system.

2. Closed-Loop Design. The open loop controller of Fig. 2b can be modified to account for the instantaneous changes in $\underline{x}$ The kth-order control sensitivity function has the general form,

$$\underline{\nu}_k = -H_{uu}^{-1}((H'_{ux} - h'_u K)\underline{\xi}_k + \underline{\beta}_k) . \tag{23}$$

Considering (19) with the remaining terms of the series taken into account, substituting for ν_k from (23) and again denote the difference between the closed loop and nominal control by $\delta u_{c\ell}$,

$$\underline{u}_{c\ell} = \underline{u}^* + \delta\underline{u}_{c\ell} \tag{24a}$$

$$\delta\underline{u}_{c\ell} = W \sum_{k=1}^{\infty} \frac{\varepsilon^k}{k!} \underline{\xi}_k + \sum_{k=1}^{\infty} \frac{\varepsilon^k}{k!} \underline{\beta}_k \tag{24b}$$

where

$$W = -H_{uu}^{-1}(H'_{ux} - h'_u K)\Big|_{\varepsilon=0} = -R_1^{-1}B'K.$$

Noting that,

$$\delta\underline{x} = \underline{x} - \underline{x}^* = \sum_{k=1}^{\infty} \frac{\varepsilon^k}{k!} \underline{\xi}_k \tag{25}$$

(24b) can be written as,

$$\delta\underline{u}_{c\ell} \cong W\delta\underline{x}_{c\ell} + \sum_{k=1}^{n} \frac{\varepsilon^k}{k!} \underline{\beta}_k. \tag{26}$$

The closed loop system after substitution of $u_{c\ell}$ will become,

$$\dot{\underline{x}}_{c\ell} = (A - BR_1^{-1}B'K)\underline{x}_{c\ell} - B \sum_{k=1}^{n} \frac{\varepsilon^k}{k!} \underline{\beta}_k + \varepsilon\underline{f}(\underline{x}_{c\ell}, u_{c\ell}) \tag{27a}$$

where

$$u_{c\ell} = -R_1^{-1}B'K\underline{x}_{c\ell} + \sum_{k=1}^{n} \frac{\varepsilon^k}{k!} \beta_k. \tag{27b}$$

Figure 2c presents the closed-loop controller for the system (3). Comparison of (19) and (26) indicates that the truncation of Taylor series for open and closed loop control is not the same. The open loop control expansion is completely truncated after n-terms while in the closed loop, only the forcing functions are truncated up to n. On the other hand, as seen from (26) the feedback term (homogeneous portion) remains exact.

If n = 0, i.e. no sensitivity functions, the open loop control of (19) is merely u* and the closed loop control of (27b) will be only the feedback signal from $\underline{x}_{c\ell}$. This rather special design is termed "zeroth-variation" system.

A summary of computational procedure of the two designs for first variation system (i.e. n=1) is given by the following steps.

1. Solve Riccati equation (15) in backward time and store matrix K and W.
2. Solve the optimal control problem (20) and (21), evaluate $\underline{\omega}_1^o$, π_1^o. Store x^*, u^*, ω_1^*, and π_1^*.
3. Solve first order adjoint system (16) in backward time and store.
4. Solve (17) for $\underline{\xi}_1$ and (18) for $\underline{\beta}_1$ and $\underline{\nu}_1$ sensitivity functions. Store $\underline{\nu}_1$, $\underline{\beta}_1$, and $\underline{\xi}_1$.
5. a) For an open loop controller solve (22).
   b) For a closed loop controller solve (27).

III. ILLUSTRATIVE EXAMPLES

The plant equation (3) indicates that the vector f may be a function of $\underline{x}$, $\underline{u}$, or both. In this section two examples are presented, one with f as a function of $\underline{u}$ only and the other a function $\underline{x}$ only. It will be also demonstrated that the parameter ε can either be chosen arbitrarily for the application of the design procedure, or be a system plant parameter whose final value depends on a given system specifications [8].

Example 1: A nonlinear plant shown in Fig. 3a is considered with the following state equations.

$$\begin{aligned} \dot{x}_1 &= -x_1 + \frac{2}{\pi}\tan^{-1}u \\ \dot{x}_2 &= x_1 - x_2 \end{aligned} \tag{28}$$

The performance index is $J = \frac{1}{2}\int_0^{4.75} (x_1^2 + x_2^2 + u^2)\,dt$, where $Q_1 = I_2$ and $R_1 = 1$. The only nonlinearity in the system is the saturation characteristic inherent in power amplifiers and common in control systems. In order to apply the method the nonlinear function of (28) is rewritten as,

$$\frac{2}{\pi}\tan^{-1}u = b_1 u + \varepsilon f_1(u)\,. \tag{29}$$

Where $b_1 = 1$, $f_1(u) = 2/\pi\tan^{-1}u - u$ and $0 \le \varepsilon \le 1$. ε is chosen as an arbitrary parameter. Considering (29) then (28) will become,

$$\begin{aligned} \dot{x}_1 &= -x_1 + u + \varepsilon f_1(u) \\ \dot{x}_2 &= x_1 - x_2\,. \end{aligned} \tag{30}$$

Note that for $\varepsilon = 0$, (30) is reduced to the "nominal" system which has a simple linear set of state equations. At $\varepsilon = 1$, (30) is equivalent to (28), the "original" system. Assuming that the states x_1 and x_2 are measurable, the nominal control is,

$$u^* = -.55x_1^* - .19x_2^*\,. \tag{31}$$

where $\underline{x}^*$ is the solution of the following homogeneous equations,

$$\begin{aligned} \dot{x}_1^* &= -1.55x_1^* - .19x_2^* \\ \dot{x}_2^* &= x_1^* - x_2^* \end{aligned} \tag{32}$$

and $\underline{x}^*(0)$ given.

Figure 4a shows the zeroth, first, and second variation closed loop control performance index versus ε for $x_1(0) = .1$, $x_2(0) = 1$. The range of the control for this set of initial conditions was found to be within the linear regions of the saturation nonlinearity (see point A, Fig. 3b). As one would expect, because the system is linear, the effect of the sensitivity analysis presented in Section II is not apparent, and there is almost very little difference between the zero-variational system, $n = 0$ and the exact optimal control law, i.e. $n = \infty$. The optimal control ($n = \infty$) was found by the "Newton's Second-Variation" method [11] which is the functional optimization version of the "Newton-Kantrovich" method in parameter optimization. As the initial conditions $\underline{x}(0)$ are increased with the control, u, nearing the knee portion of the saturation

curve (see point B in Fig. 3b), the new J vs. $\mathcal{E}$ plots, shown in Fig. 4b reveal the effectiveness of the design. As seen there is an improvement on J from n= 0 to n= 1 and 2. The second variation performance index at $\mathcal{E}$= 1 seems to indicate that this control is "close enough" to the optimal one so that the computational difficulties (computer time, nonlinearities, etc.) of the latter may be omitted. Figure 4c presents a comparison between open and closed loop performances for $\underline{x}(0)$ well in the nonlinear region of $f_1(u)$. It is apparent that with almost the same amount of computation an improvement has been made in performance of the system by adopting a closed loop controller.

Example 2. This example considers the synthesis of a "control logic for a regulation system" [4] whose torque source is a field-controlled d.c. motor. The state equations pertaining to this system are,

$$J_L \frac{d^2(\delta\theta)}{dt^2} + B_L \frac{d(\delta\theta)}{dt} = (K_i E/R_a) i_f - (K_i K_b/R_a) \frac{d(\delta\theta)}{dt} i_f^2$$
$$L_f \frac{di_f}{dt} + R_f i_f = e_f. \qquad (33)$$

In (33), J_L = moment of inertia of the load, $\delta\theta$ = angular pos. error of the shaft, B_L = friction loss coefficient, $\delta\theta = \delta\omega$ is the angular velocity error of the shaft, K_i = torque constant, L_f = field inductance, R_f = field resistance, e_f = field voltage, E = constant amature voltage, R_a = armature resistance, and K_b = back emf constant. Armature inductance has been neglected. The nonlinear term in (33) arises from the magnetization properties of the motor's ferromagnetic material. Equation (33) can be rewritten as,

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = a_{22}x_2 + a_{23}x_3 - \mathcal{E}x_2x_3^2 \qquad (34)$$
$$\dot{x}_3 = a_{33}x_3 + b_3u$$

where $x_1 = \delta\theta$, $x_2 = \delta\omega$, $x_3 = i_f$, $u = e_f$, $a_{22} = -B_L/J_L$, $a_{23} = K_iE/J_LR_a$, $a_{33} = -R_f/L_f$, $b_3 = 1/L_f$, and $= K_iK_b/J_LR_a$. Since R_a has relatively large value, the magnitude of $\mathcal{E}$ is small compared to a_{22} and a_{23}, and it is normally set equal to zero and the nonlinear term is neglected [4]. The performance index is $J = \frac{1}{2}\int_0^\infty (\underline{x}'Q\underline{x} + r_1 \cdot u^2)dt$. The control problem is to find a suitable near-optimal control which approximately minimizes J. Both open and closed loop controllers for n= 0 and n= 1 cases are considered. For computational purposes the same numerical values are used as in [4] so that this controller results can be compared with that developed by Garrard, et. al.

$$\dot{\underline{x}} = A\underline{x} + Bu + \mathcal{E}\underline{f}(\underline{x}) \qquad (35a)$$

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -.2 & 2 \\ 0 & 0 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad f = \begin{bmatrix} 0 \\ -x_2x_3^2 \\ 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad r_1 = 1, \qquad (35b)$$

and $\mathcal{E}_f$ = .20. The nominal control and state equations are,

$$u^* = -R_1^{-1}B'KX^* = -x_1^* - 2.12x_2^* - .872x_3^* \qquad (36a)$$

$$\begin{aligned} \dot{x}_1^* &= x_2^* \\ \dot{x}_2^* &= -.2x_2^* + 2x_3^* \\ \dot{x}_3^* &= -x_1^* - 2.12x_2^* - 5.872x_3^* \,. \end{aligned} \qquad (36b)$$

Table 1 summarizes the performance idex J for the four designs considered and the ones by the method developed in [4]. The first variation closed loop design gives the best result as anticipated. There is a clear improvement on the method developed by Garrard, et.al.

Table 1

control law		Performance Index, J Initial Conditions, $\underline{X}(0)$		
		$\begin{bmatrix} .10 \\ .10 \\ .10 \end{bmatrix}$	$\begin{bmatrix} .25 \\ .25 \\ .10 \end{bmatrix}$	$\begin{bmatrix} .50 \\ .50 \\ .10 \end{bmatrix}$
Open loop	n=0	0.2294	1.1330	4.0508
	n=1	0.2214	1.0951	4.0166
Closed loop	n=0	0.2165	1.0724	3.9520
	n=1	0.2162	1.0712	3.9349
Method developed in [4]		0.2288	1.1259	4.0627

Figure 5a shows the angular position error versus time for the four control laws considered. The open loop designs (in particular n= 0 which has u* as its control) are clearly unsatisfactory due to large steady state errors. The closed loop first variation design for $\mathcal{E}$ = .2 resulted the lowest error in θ. The angular velocity error versus time plots are presented in Fig. 5b. This figure indicates that all three controls have sufficiently small steady state error in angular velocity. The field voltage for two of the designs are shown in Fig. 5c. As seen the closed loop control has a lower demand for field voltage. Furthermore the closed loop controller of this paper allows the designer of such a system to study the effects of varying the plant parameter $\mathcal{E}$ under more realistic conditions.

IV. CONCLUSIONS

The design procedure introduced in this paper approximates the performance index up to 2n+1 terms of Taylor series expansion as in the optimally adaptive control.† This controller also makes use of the convenient derivation of sensitivity equations as in the optimally sensitive control and extended it to nonlinear systems. For each new approximation, the sensitivity functions $\underline{\nu}_{i+1}$ and $\underline{\beta}_{i+1}$ are computed from

†See reference [5] for a proof of this assertion by Werner and Cruz.

the previous functions $\underline{\nu}_i$, $\underline{\xi}_i$, $\underline{\alpha}_i$, $\underline{\beta}_i$, the current solutions of $\underline{\xi}_{i+1}$, and adjoint system $\underline{g}_{i+1}$. Therefore, only the storage of the i^{th} order is needed for the $i+1^{st}$.

The main disadvantage of the design is the number of partial differentiations needed for finding each new approximated controller. For a "third variation" system, one has to find second partials of ω_1, π_1, θ_1, F, Q, S, and T with respect to $\mathcal{E}$ and f with respect to $\underline{u}$ and $\underline{x}$. This difficulty may be eliminated by checking the performance and responses of each new design with the previous one and if the last approximation does not improve performance by a "great deal", the search for a more optimal design may terminate. This was demonstrated in the first example.

The choice between an open and closed loop design is based on two contrasting factors, one is the system response characteristic and the other is the cost of constructing the controller. The designer may prefer a high-order open loop controller with poorer response and performance characteristics over a lower order closed loop which is costier to implement.

ACKNOWLEDGMENTS

The author wishes to extend his appreciation to Professors J. B. Cruz, Jr. and P. V. Kokotović whose directions and criticisms made writing this paper possible. Also, to the Coordinated Science Laboratory for their financial support.

REFERENCES

1. Pearson, J. D., J. Electronic Control, Vol. 13, pp. 453-469, 1960.

2. Durbeck, R. C., IEEE Transactions on Automatic Control, Vol. AC-10, pp. 144-149, 1965.

3. Al'Brekht, E. G., Journal Applied Math. Mechanics, Vol. 25, pp. 1252-1266, 1961.

4. Garrard, W. L., N. H. McClamroch, and L. G. Clark, "An Approach to Suboptimal Feedback Control of Nonlinear Systems," International Journal of Control, Vol. 5, No. 5, 1967, pp. 425-35.

5. Werner, R. A. and J. B. Cruz, Jr., "Feedback Control which Preserves Optimality for Systems with Unknown Parameters," IEEE Trans. on Automatic Control, AC-13, No. 6, December, 1968.

6. Kelly, H. J., "An Optimal Guidance Approximation Theory," IEEE Trans. on Automatic Control, Vol. AC-9, No. 4, October 1964, pp. 375-80.

7. Kokotović, P. V., J. B. Cruz, Jr., J. E. Heller, and P. Sannuti, "Synthesis of Optimally Sensitive Systems," Proceedings of the IEEE, Vol. 56, No. 8, August 1968, pp. 1318-24.

8. Jamshidi, M. S., "A New Design Technique of Suboptimal Nonlinear Automatic Feedback Control Systems Using Sensitivity Functions," Report R-405, Coordinated Science Laboratory, Univ. of Ill., Urbana, Ill., January 1969.

9. Kalman, R. E., "Contributions to Theory of Optimal Control," Bol. Soc. Math. (Mexico), 2nd Ser., Vol. 5, pp. 102-118, 1960.

10. Athans, Michael and Peter L. Falb, Optimal Control, McGraw-Hill Book Co., New York, 1966, pp. 767-773, 793-812.

11. Eveleigh, Virgil W., Adaptive Control and Optimization Techniques, McGraw-Hill Book Co., New York, 1965, pp. 38-45.

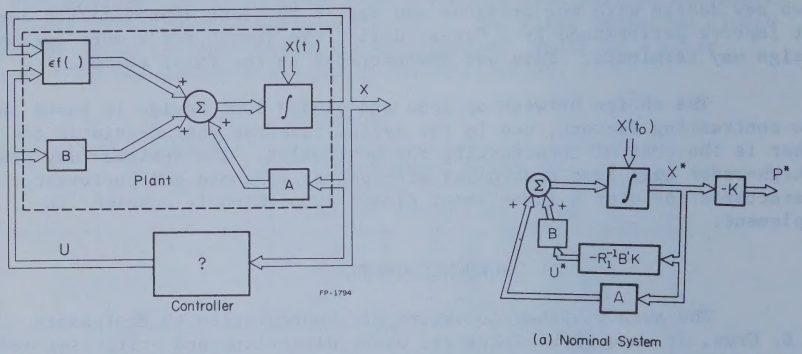


Fig. 1 Control Problem

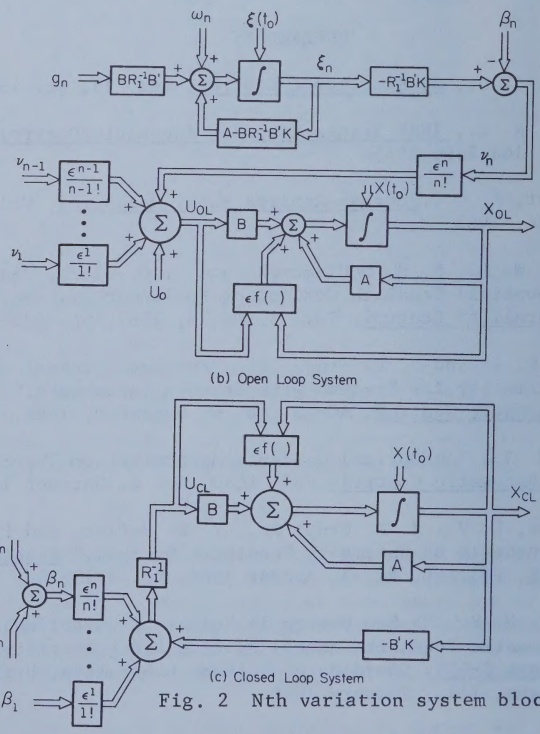
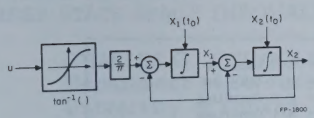
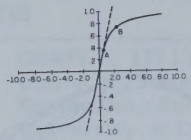


Fig. 2 Nth variation system block diagram

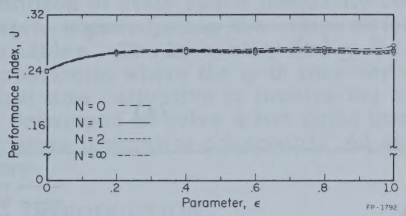


(a)

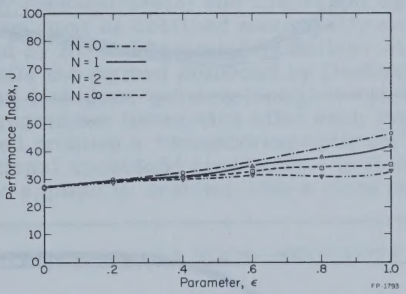


(b)

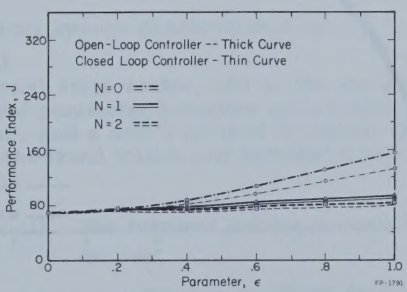
Fig. 3 Example 1 a simple nonlinear plant



(a) $x_1^o = .1,\ x_2^o = 1.0$

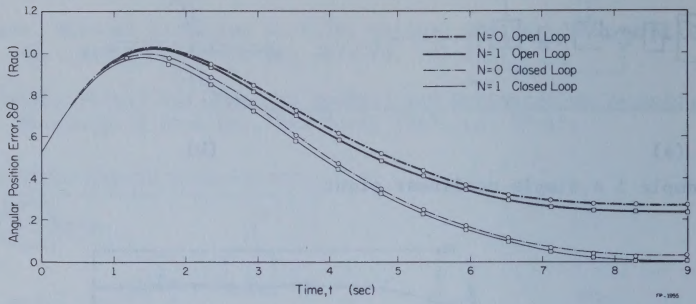


(b) $x_1^o = 10.,\ x_2^o = .10$

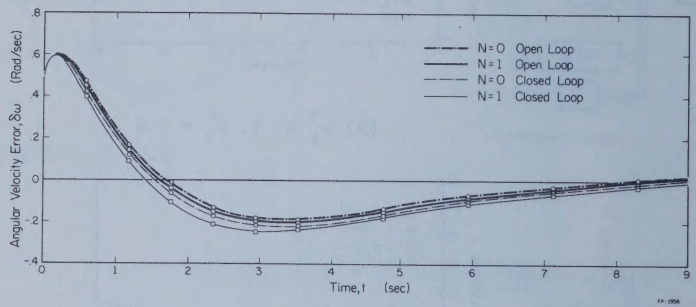


(c) $x_1^o = 10.,\ x_2^o = 10.$

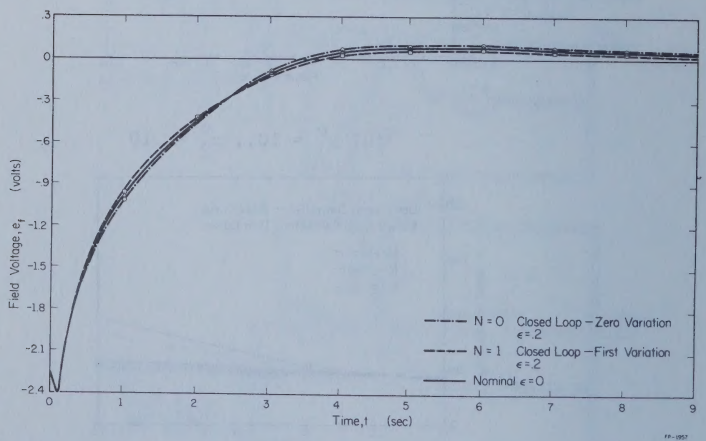
Fig. 4 Performance Index vs. $\mathcal{E}$



(a)



(b)



(c)

Fig. 5 Example 2 responses

OPTIMAL SOLUTION OF TWO POINT BOUNDARY VALUE PROBLEMS WITH Q-TH ORDER STATE SPACE INEQUALITY CONSTRAINTS

CHARLES C. SCHOOLER
Department of Electrical Engineering
University of Kansas
Lawrence, Kansas

GERRY L. KELLY
Center for Research Inc.
Engineering Science Division
University of Kansas
Lawrence, Kansas

ABSTRACT

The problem considered in this paper is the Mayer formulation of the problem of Bolza with the addition of state space inequality constraints. The state inequality constraint, imposed along the entire trajectory, is a function $S(x)$ of the state variables. The inequality $S(x) \leq 0$ is a general q-th order state inequality constraint where the q-th time derivative of the constraint function is the first time derivative to involve the control variable explicitly. An algorithm is developed to solve a two point boundary value problem having a q-th order state inequality constraint. An example is included to illustrate the theory.

INTRODUCTION

The solution of differential equations in control problems with an inequality constraint on the state and/or control variables was first considered by Valentine [1] and later by Gamkrelidze[2] and Dreyfus[3] as well as by others. In general a solution must be obtained numerically and numerical methods have been proposed by Ho and Brentani[4], Kelley, et.al.[5], and Denham and Bryson[6]. While the method proposed by Denham and Bryson is perhaps the most powerful technique yet developed, it can become rather involved and may require programmer interaction after each iteration.

For the unconstrained problem a "neighboring optimum" technique has the advantages of computational speed and simplicity. In this paper it is shown the technique may be applied to systems with a state inequality constraint.

OPTIMAL PROBLEM FORMULATION WITH A STATE INEQUALITY CONSTRAINT

The Mayer formulation of the problem of Bolza is defined and the necessary conditions for optimality are given. The salient features of these conditions are discussed and comparisons are made with the unconstrained case.

Consider the dynamical system characterized by

$$\dot{x} = f(x(t), u(t), t) \tag{1}$$

where $x(t)$ is the n-dimensional state vector, $u(t)$ is the scalar* control variable, t is the independent parameter hereafter called time, f is an n-vector of specified functions, and a dot($\dot{}$) is used to denote time differentiation. Let t_o and $T > t_o$ be the fixed initial and terminal times, respectively. Then

$$x_o \triangleq x(t_o) \tag{2}$$

is the given initial state for (1). The terminal boundary condition is assumed to be a point given by

*Since the scalar constraint as an equality can in effect determine only one component of the control vector, the usual vector control is reduced to a scalar to avoid having one component determined by one relationship and the remaining components by another relationship.

$$x_T = x(T) \tag{3}$$

A scalar valued function describing the system performance is given by

$$J = \Phi[x(T), T] \tag{4}$$

In addition the system must satisfy the state space inequality constraint described by the scalar function $S(x)$:

$$S(x) \leq 0 \tag{5}$$

The state inequality constraint is the general q-th order form ($q \leq n$) so that

$$\begin{aligned} &\partial/\partial u\ (d^j S(x)/dt) = 0 && j = 0, 1, \ldots, q-1 \\ &\partial/\partial u\ (d^j S(x)/dt) \neq 0 && j = q \end{aligned} \tag{6}$$

The state vector x is assumed to be continuous on the interval (t_o, T) and its time derivative is assumed to be continuous almost everywhere on (t_o, T). The trajectory is comprised of interior subarcs ($S(x) < 0$) and boundary subarcs ($S(x) = 0$). It is assumed the optimal trajectory contains at most one boundary subarc since a problem with multiple boundary subarcs may be considered as a series of problems with at most one boundary subarc. Define $t_1 > t_o$ as the time eq. (5) first becomes an equality. Similarly define $t_2 > t_1$ as the time eq. (5) again becomes a strict inequality. The state inequality relationship $S(x)$ and its first $q-1$ time derivatives are assumed to be continuous almost everywhere on (t_o, T) and in addition must be continuous in a neighborhood of t_1.

The necessary conditions for a local extremal have been derived from the calculus of variations. Following the usual approach in the calculus of variations the system equations and constraints are appended to the cost function. However, the general q-th order state inequality constraint may not be directly appended. Since the q-th time derivative of $S(x)$ is the first time derivative to explicitly involve the control u, the constraint function may be considered as an equality constraint on the q-th time derivative over the interval (t_1, t_2) and is given by

$$S^{(q)}(x) \triangleq d^q S(x)/dt^q = 0 \tag{7}$$

To insure that the constraint function $S(x)=0$ on the boundary it is necessary that $N(x(t_1), t_1)$ be the zero vector where

$$N^T(x(t_1), t_1) \triangleq (S(x), \dot{S}(x), \ldots, S^{q-1}(x)) \tag{8}$$

The q-vector N may be regarded as an "intermediate" boundary constraint. The necessary conditions for local optimality are then, in addition to the system equations,

$$\dot{\lambda}^T(t) = -\lambda^T(t)\ (\partial f/\partial x) \quad : S(x) < 0 \tag{9}$$

$$\dot{\lambda}^T(t) = -\lambda^T(t)[(\partial f/\partial x) - (\partial f/\partial u)(\partial S^{(q)}/\partial u)^{-1}(\partial S^{(q)}/\partial x)] : S(x) = 0 \tag{10}$$

where $\lambda(t)$ is the n-dimensional adjoint variable.

Along an interior subarc the optimal control is determined from

$$\lambda^T(t)(\partial f/\partial u) = 0 \tag{11}$$

At the entry corner defined at $t = t_1$,

$$\lambda^T|_{t_1^-} = \lambda^T|_{t_1^+} + \gamma^T(\partial N/\partial x)|_{t_1} \tag{12}$$

$$\lambda^T\dot{x}|_{t_1^-} = \lambda^T\dot{x}|_{t_1^+} - \gamma^T(\partial N/\partial t)|_{t_1} \tag{13}$$

where, in anticipation of a discontinuity at the entry point, $t_1^- = t_1 - \epsilon$ and $t_1^+ = t_1 + \epsilon$ where ϵ is an arbitrarily small positive constant. The q-dimensional vector γ is a time invariant Lagrange multiplier. Similarly, at the exit corner defined at $t = t_2$,

$$\lambda^T|_{t_2^-} = \lambda^T|_{t_2^+} \tag{14}$$

$$\lambda^T\dot{x}|_{t_2^-} = \lambda^T\dot{x}|_{t_2^+} \tag{15}$$

Equations (9) and (11), in addition to the system equations (1-3), are the necessary conditions for extremalizing (4) when there is no state inequality constraint. With the imposition of a state inequality constraint, the control along a boundary subarc is determined from (7) and the form of the differential system describing λ is altered as indicated by (10). Moreover, at the entry point, λ and the Hamiltonian function $H \triangleq \lambda^T \dot{x}$ are in general discontinuous. At the exit point, however, both are assumed continuous. It is possible to have discontinuities at the exit point as well, but no generality is lost by assuming continuity and computational advantages are gained. [7]

NUMERICAL SOLUTION: A NEIGHBORING OPTIMUM TECHNIQUE

Equations (1) and (9) or (10) together comprise a canonical differential system of order 2n. With n boundary conditions specified at $t=t_o$ by x_o and n boundary conditions specified at $t=T$ by x_T, the solution of this problem cannot in general be obtained in closed form and an iterative technique must be employed. The most direct way of solving this problem is to assume n additional boundary conditions at either boundary. Suppose the initial boundary defined at $t=t_o$ is selected. Then an initial estimate of the adjoint variable

$$\lambda_o \triangleq \lambda(t_o) \tag{16}$$

is made. There are then 2n boundary conditions defined at $t=t_o$. With a full complement of boundary conditions at $t=t_o$ the solution of the canonical differential system may be determined over the interval (t_o, t_1). At t_1, however, λ is in general discontinuous and the magnitude of the discontinuity cannot be predicted since the q-vector γ is unknown. Moreover the entry time t_1 is also unknown. If γ were known, the discontinuity in λ could be predicted from (12) and t_1 could be determined from (13).* Equation (13) serves as a transversality condition to define t_1. Having calculated the discontinuity at $t=t_1$ the solution of the canonical system described by (1) and (10) may be determined on the interval (t_1, t_2) where t_2 is determined from (15). By (14) and the continuity assumptions on x, the solution of the canonical system is continuous at $t=t_2$. The canonical differential system described by (1) and (9) may then be integrated over the interval (t_2, T).

To perform the integration of the canonical differential system, a total of n+q estimates had to be made.

$$c \triangleq (\lambda_1(0), \ldots, \lambda_n(0), \gamma_1, \ldots, \gamma_q) \tag{17}$$

To correct these n+q estimates, a total of n+q comparisons with the optimal trajectory are needed. These are supplied by (3) and (8).

$$y \triangleq (x_1(T), \ldots, x_n(T), S(x), \ldots, S^{(q+1)}(x)) \tag{18}$$

At this point some correction procedure must be used, and the procedure chosen is the Newton-Raphson iterative technique. Its most important feature is its rapid convergence to the solution, but convergence is guaranteed only if the initial estimates are sufficiently close to the optimal solution. Sufficient conditions for convergence as well as an estimate of the rate of convergence have been established by Kantorovich. [8,9]

The Newton-Raphson successive approximation technique may then be represented as

$$c^{(k+1)} = c^{(k)} - \left(\partial y(c)/\partial c\right)^{-1}\Big|_{c^{(k)}} \left(y^{(c)} - y^*\right)\Big|_{c^{(k)}} \tag{19}$$

where the superscript k denotes the index of approximation and the asterick (*) denotes the optimal value of (18).

*In the event (13) is satisified only on the optimal trajectory, t_1 must be estimated along with λ and γ. All terms of (13) are then grouped together and considered as the q+1 term on N in eq. (8).

In the event y as a function of c exists only implicitly it is necessary to employ numerical methods to determine the matrix of partial derivatives in (19). An accurate estimate of these partials may be obtained by solving the perturbation equation of the canonical system.[8]

A "neighboring optimum" solution of the optimal control problem may now be summarized in the following steps:

1. An initial estimate is made of λ_o and γ.
2. The differential equations (1) and (9) are intergrated using the control determined from (11) for all $t < t_1$ where t_1 is determined from (13).
3. The discontinuity in λ is computed from (12). Since x is absolutely continuous, both λ and x are defined at $t = t_1^+$. The vector N is evaluated and, except on an optimal trajectory, will be nonzero.
4. The differential equations (1) and (10) are integrated using the control determined from (7) for $t_1 < t \leq t_2$ where t_2 is determined from (15).
5. By (14) λ is continuous at $t = t_2$. Equations (1) and (9) are then integrated using (11) for $t_2 \leq t \leq T$.
6. The resulting $x(T)$ is computed and the difference with (3) determined. The initial estimates in step 1 are then corrected using (19).
7. Steps 2-6 are repeated until N is sufficiently close to zero and the computed x(T) sufficiently close to the desired terminal condition.

AN EXAMPLE

In this section a simple second order example is presented to illustrate the numerical techniques. The system is a linear time invariant system described by

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned} \tag{20}$$

The initial and final times are given by $t_o = 0$ and $T = 1$. The initial and terminal states are

$$x^T(0) = (0,1) \qquad x^T(T) = (0,0) \tag{21}$$

The Mayer formulation of an integral cost function is assumed to be

$$J = x_3(T) \tag{22}$$

where

$$\dot{x}_3 = u^2/2 \tag{23}$$

The state inequality constraint is given by

$$S(x) = .25 - x_2 \leq 0 \tag{24}$$

The differential equations (9) and (10) describing the adjoint variable are identical.

$$\dot{\lambda}^T = (0, -\lambda_1, 0) \tag{25}$$

The vector N has a single component given by $(.25 - x_2)$. The corner conditions of (12) and (13) then become

$$(\lambda_1(t_1^-), \lambda_2(t_1^-)) = (\lambda_1(t_1^+), \lambda_2(t_1^+) - \gamma) \tag{26}$$

$$\lambda_2 u \,|\, t_1^- = \lambda_2 u \,|\, t_1^+ \tag{27}$$

The components x_3 and λ_3 have been omitted since they do not enter into the computation.

Initial estimates of $\lambda_1(0)$, $\lambda_2(0)$, and γ were made. Rather than use eq. (27) to determine t_1, the entry time was also estimated and (27) was considered as an additional component of N. The following set of initial estimates required 18 iterations to converge to five decimal place accuracy.

$$\begin{bmatrix} \lambda_1(0) \\ \lambda_2(0) \\ \gamma \\ t_1 \end{bmatrix} = \begin{bmatrix} 8.0 \\ 5.0 \\ -6.0 \\ .5 \end{bmatrix} \xrightarrow[\text{iterations}]{18} \begin{bmatrix} 8.2423 \\ 4.5393 \\ -6.5691 \\ .55074 \end{bmatrix}$$

CONCLUSIONS

This paper deals with the fixed terminal time problem whose terminal constraint is a point in the n dimensional vector space of x. An extension to free terminal time problems with an m-dimensional ($m < n$) terminal manifold will alter the details but not the substance of this paper. Moreover, as exemplified in the example it is possible to estimate more than (n+ q) parameters.

As indicated by the relative scarcity of numerical techniques for constrained optimization problems, investigators have been reluctant to examine problems with general q-th order state inequality constraints. In this paper we have attempted to show that, at least conceptually, the constrained problem may be treated as an unconstrained problem with "intermediate" boundary constraints.

ACKNOWLEDGEMENT

The use of the Computation Center facilities at the University of Kansas is gratefully acknowledged.

REFERENCES

1. Valentine, F. A., "The problems of Lagrange with differential inequalities as added side conditions," in Contributions to the Theory of Calculus of Variations, 1933-37, University of Chicago Press, Chicago, 1937, pp. 405-447.
2. Gamkrelidze, R. V., "Optimal processes with restricted phase coordinates," Izv. Akad. Navk SSSR, Ser. Mat., 1960, Vol. 24, pp. 315-356.
3. Dreyfus, S., "Variational problems with inequality constraints," J. of Mathematical Analysis and Applications, 1962, Vol. 4, pp. 297-308.
4. Ho, Y. C. and P. B. Brentani, "On computing optimal control with inequality constraints," J. Siam Control , Ser. A., 1963, Vol. 1 pp. 319-348.
5. Kelley, H. J., M. Falco, and D. J. Ball, "Air vehicle flight path optimization," Vol. 1, Rept. AFOSR/DRA-62-4, Air Force Office of Sci. Res., Holloman AFB, New Mexico, 1962.
6. Denham, W. F. and A. E. Bryson, "Optimal programming problems with inequality constraints II: solution by steepest ascent," AIAA Journal, 1964, Vol. 2, pp. 25-34.
7. Bryson , A. E., W. F. Denham, and S. Dreyfus, "Optimal programming problems with inequality constraints I: necessary conditions for optimality," AIAA Journal, 1963, Vol. 1, pp. 2544-2550.
8. Payne, J. A., "Computational methods in optimal control," Ph.D. Dissertation, University of California, Los Angeles, 1965.
9. Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, New York, John Wiley, 1962, pp. 366-374.
10. McIntyre, J., and B. Paiewonsky, "On optimal control with bounded state variables," in Advances in Control Systems, Vol. 5, C.T. Leondes, editor, Academic Press, New York, 1967, pp. 389-419.

A STUDY OF THE OUTPUT REGULATOR PROBLEM FOR LINEAR SYSTEMS WITH INPUT VECTOR

UMBERTO DI CAPRIO and PAUL P. WANG
Department of Electrical Engineering
Duke University
Durham, North Carolina 27706

The difficult problem of synthesizing an optimal regulation scheme for linear time-invariant systems, assuming an integral quadratic cost functional, when the desired output is a constant non-zero vector and when the upper time limit approaches infinity is studied in this paper. The problem is solved by properly redefining the cost functional so that consistency with the stability requirement is achieved. The resultant structure of the optimal system is also included.

Given the linear, time invariant, completely controllable and observable system

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{B}\,u \tag{1}$$

$$\underline{y} = \underline{C}\,\underline{x} \tag{2}$$

where

$\underline{x}$: state vector having the dimensions (nx1)
$\underline{u}$: control vector having the dimensions (rx1)
$\underline{y}$: output vector having the dimensions (mx1)

and

$$0 < m \leq r \leq n$$

and given the cost functional

$$J = \frac{1}{2}\int_0^\infty \left[< \underline{y}(t), \underline{Q}\,\underline{y}(t) > + < \underline{u}(t), \underline{R}\,\underline{u}(t) >\right]dt \tag{3}$$

where $\underline{u}(t)$ is not constrained and $\underline{Q}$, $\underline{R}$ are constant, positive definite matrices, then the unique optimal control law is [1]

$$\underline{u}(t) = -\underline{R}^{-1}\,\underline{B}^T\,\underline{K}\,\underline{x}(t) \tag{4}$$

where $\underline{K}$ is a (nxn) constant positive definite matrix. It should be noted that the constant nonlinear Riccati Matrix Algebraic equation is also satisfied:

$$-\underline{K}\,\underline{A} - \underline{A}^T\,\underline{K} + \underline{K}\,\underline{B}\,\underline{R}^{-1}\,\underline{B}^T\,\underline{K} - \underline{C}^T\,\underline{Q}\,\underline{C} = \underline{0} \tag{5}$$

It follows from (4) that

$$\dot{\underline{x}} = \left(\underline{A} - \underline{B}\,\underline{R}^{-1}\,\underline{B}^T\,\underline{K}\right)\underline{x} \quad . \tag{6}$$

The above system is homogeneous with asymptotically stable solutions. Therefore $\lim_{t\to\infty} \underline{x}(t) = \underline{0}$. It obviously follows that $\lim_{t\to\infty} \underline{y}(t) = \underline{0}$ and $\lim_{t\to\infty} \underline{u}(t) = \underline{0}$.

An obvious but extremely important and significant generalization of the above mentioned problem requires that the optimal control law (if any exists) keep the $\underline{y}(t)$ near some desired constant non-zero vector $\underline{z}_d$. This problem might be considered as a specific tracking problem. (It is designated here as the "non-trivial output regulator problem") The theory now available has not yet [1] given a satisfactory analytical solution to the non-trivial output regulator problem when the assumed

cost functional is

$$J_1 = \frac{1}{2}\int_0^\infty \left[< \underline{e}(t), \underline{Q}\,\underline{e}(t) > + < \underline{u}(t), \underline{Ru}(t) > \right] dt \qquad (7)$$

where

$$\underline{e}(t) = \underline{y}(t) - \underline{z}_d, \quad \underline{z}_d \neq \underline{0} .$$

The difficulty resides in the "bad" choice of J_1 because the existence of a (finite) minimum J_1 requires that

$$\lim_{t\to\infty} \underline{e}(t) = \underline{0}, \text{ or } \lim_{t\to\infty} \underline{y}(t) = \underline{z}_d \neq \underline{0} \text{ and } \lim_{t\to\infty} \underline{u}(t) = \underline{0},$$

which implies that the optimal system should have a stable equilibrium state $\underline{x}_d \neq \underline{0}$ such that $\underline{C}\,\underline{x}_d = \underline{z}_d \neq \underline{0}$. This is possible if the optimal control law contains a constant component $\underline{u}_d$ such that

$$\underline{A}\,\underline{x}_d + \underline{B}\,\underline{u}_d = \underline{0} . \qquad (8)$$

Now, we claim that it is always possible to reduce the non-trivial output regulator problem to a (trivial) output regulator problem, provided that the cost functional is defined as:

$$J_2 = \frac{1}{2}\int_0^\infty \left[< \underline{e}(t), \underline{Q}\,\underline{e}(t) > + < \left(\underline{u}(t) - \underline{u}_d\right), \underline{R}\left(\underline{u}(t) - \underline{u}_d\right) > \right] dt. \qquad (9)$$

Then the optimal control law becomes

$$\underline{u}(t) = -\underline{R}^{-1}\,\underline{B}^T\,\underline{K}\left(\underline{x} - \underline{x}_d\right) + \underline{u}_d .$$

The resultant optimal closed loop system yields

$$\dot{\underline{x}} = \left(\underline{A} - \underline{B}\,\underline{R}^{-1}\,\underline{B}^T\,\underline{K}\right)\underline{x} + \underline{B}\,\underline{r}_d = \underline{G}\,\underline{x} + \underline{B}\,\underline{r}_d$$

where

$$\underline{r}_d = \underline{u}_d + \underline{R}^{-1}\,\underline{B}^T\,\underline{K}\,\underline{x}_d .$$

The structure of the optimal non-trivial output regulator is shown in Figure 1.

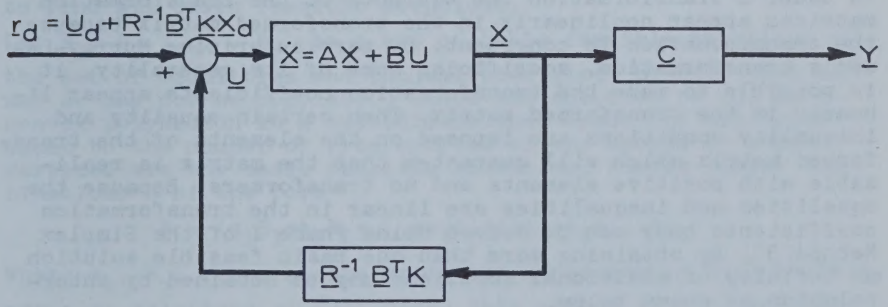


Figure 1. Optimal closed loop scheme for the non-trivial output regulator problem.

Reference

[1] Athans, M. and Falb, P. L., Optimal Control, McGraw Hill, Inc., New York, N. Y., 1966, pp. 750-814.

LINEAR PROGRAMMING APPLIED TO THE COMPUTER GENERATION OF EQUIVALENT CIRCUITS

I. D. RAPPORT
Case Western Reserve Univ.
Cleveland, Ohio

M. A. MURRAY-LASSO
Universidad Nacional Autónoma de México.
México, D.F., MEXICO

ABSTRACT

A method for automatically obtaining equivalent grounded circuits with positive elements by starting with arbitrary Y matrices whose inverse realizes one or several driving-point or transfer impedances is given. The method is based on using a single-sided transformation and solving a set of simultaneous linear equalities and inequalities with linear programming techniques. An example is given.

INTRODUCTION

Although a very general method of obtaining equivalent circuits by linear transformations was proposed by Cauer[1] in 1929, in practice it is difficult to apply the transformation and be sure beforehand that it will yield circuits with positive elements requiring no transformers. The reason is that the elements of the transformed matrix must satisfy certain inequality conditions that are difficult to control. Murray-Lasso[2] has generalized Cauer's transformation to be able to deal conveniently with active, non-bilateral circuits. The resulting transformation is not necessarily congruent. No indication was given by Murray-Lasso of how to choose the transformation matrices to yield positive valued elements. This paper is devoted to this question. A solution is given based on the application of linear programming techniques.

LINEAR PROGRAMMING APPROACH

In Cauer's transformation the elements of the transformation matrices appear nonlinearly in the transformed matrix because the transformation is congruent. By particularizing Murray-Lasso's transformation, sacrificing some of its generality, it is possible to make the transformation coefficients appear linearly in the transformed matrix. Then certain equality and inequality conditions are imposed on the elements of the transformed matrix which will guarantee that the matrix is realizable with positive elements and no transformers. Because the equalities and inequalities are linear in the transformation coefficients they can be solved using Phase I of the Simplex Method[3]. By obtaining more than one basic feasible solution an infinity of additional solutions may be obtained by interpolation as shown below.

PROBLEM FORMULATION

Following Murray-Lasso[2], let the nodal equations of a circuit be

$$YV = I \tag{1}$$

where Y is an nxn nodal admittance matrix, V is an n-vector of node voltages, and I is an n-vector of source-injected node currents. After nonsingular transformations $V = A\hat{V}$ and

$I = B\hat{I}$, the transformed nodal admittance matrix is

$$\hat{Y} = A^{-1}YB \tag{2}$$

To keep ceratain circuit functions invariant the matrices A^{-1} and B are restricted to have certain forms. Let us now choose the entries of the matrix A^{-1} within the allowable form (for example $A^{-1} = U$, where U is a unit matrix) and also choose the form of B according to the particular objective but leaving the free entries of B as unknowns. The elements of $\hat{Y}$ will be linear combinations of the free elements b_{ij} of the matrix B. Assume for simplicity that the circuit in question is a transformerless positive valued resistive circuit. The restrictions

$$\begin{aligned} &\hat{g}_{ii} \geq 0 \ ; \ i=1,2,\ldots,n \\ &\hat{g}_{ij} \leq 0 \ ; \ i\neq j \ ; \ i,j=1,2,\ldots,n \\ &\hat{g}_{ij} = g_{ji} \ ; \ i\neq j \ ; \ i,j=1,2,\ldots,n \\ &\sum_{j=1}^{n} \hat{g}_{ij} \geq 0 \ ; \ i=1,2,\ldots,n \end{aligned} \tag{3}$$

which are linear in the elements of the matrix B, guarantee that the nodal admittance matrix $\hat{Y}$ is realizable with positive resistors without transformers. The linear inequalities and equalities (3) may be solved for the unkown free elements of B using the same techniques that are emplyed to solve linear programming problems where equality and inequality constraits appear. In particular Phase I of the Simplex Method[3] can be used to find a first feasible solution. (Note that we are not trying to optimize a cost function but only trying to solve a set of linear inequalities and equalities. Thus zero cost coefficients may be used with any available linear programming routine.) Additional feasible solutions may be obtained at a very small computational cost by using the maze solving method proposed by Charnes[4]. With this method it is possible to find all basic feasible solutions.

Let B_1, B_2, ... , B_f be a set of basic feasible solutions of the set of equalities and inequalities given by (3). (The set may or may not be complete.) Since the solution of (3) is a convex polyhedron in a multidimensional space (of dimension equal to the number of free entries of the matrix B) whose vertices are the basic feasible solutions, convex linear combinations B of the form

$$B = (a_1B_1+a_2B_2+\ldots+a_fB_f)/(a_1+a_2+\ldots+a_f)$$

where a_1, a_2, ... , a_f are arbitrary non-negative real numbers, are also solutions of (3). Using this method of interpolation of B's an infinity of transformation matrices all of which are guaranteed to yield equivalent transformerless positive-valued resistive circuits may be obtained very simply.

SIMPLE ILLUSTRATIVE EXAMPLE

Suppose it is desired to realize a transfer resistance of 0.2 ohms. A candidate Y matrix that realizes it is

$$Y = \begin{bmatrix} -11/3 & 3 & -10/3 \\ -1/3 & 5 & 1/3 \\ 13/3 & -8 & 8/3 \end{bmatrix} \qquad (4)$$

A realization of Y is shown in Fig. 1 . Obviously the circuit cannot be realized with positive resistors. Since Y is not symmetric it is not even realizable with bilateral elements. (Matrices like Y of (4) are very easy to obtain. For this example all that need be done is to start with any nodal impedance matrix $Z = (z_{ij})$ with $z_{21} = 0.2$ and invert it to get a valid Y).

The relations corresponding to (3) when written in matrix form are:

$$\left[\begin{array}{cccccc} -11/3 & 0 & 0 & -10/3 & 0 & 0 \\ 0 & 0 & 13/3 & 0 & 0 & 8/3 \\ 13/3 & 13/3 & 13/3 & 8/3 & 8/3 & 8/3 \\ \hline -1/3 & 11/3 & 0 & 1/3 & 10/3 & 0 \\ 13/3 & 0 & 11/3 & 8/3 & 0 & 10/3 \\ 0 & 13/3 & 1/3 & 0 & 8/3 & -1/3 \\ \hline 0 & 0 & -11/3 & 0 & 0 & -10/3 \\ 0 & 0 & -1/3 & 0 & 0 & 1/3 \\ -1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & -1/3 & 0 \\ 11/3 & 11/3 & 11/3 & 10/3 & 10/3 & 10/3 \\ 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 \end{array}\right] \left[\begin{array}{c} b_{11} \\ b_{12} \\ b_{13} \\ b_{31} \\ b_{32} \\ b_{33} \end{array}\right] \begin{array}{c} \\ \geq \\ \\ = \\ \\ \leq \\ \\ \end{array} \left[\begin{array}{c} 0 \\ 0 \\ 8 \\ \hline 3 \\ 0 \\ 8 \\ \hline 0 \\ 0 \\ 0 \\ 5 \\ 3 \\ 5 \end{array}\right] \qquad (5)$$

where the b_{ij} are the elements of the transformation matrix B of (2) which has the form

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 0 & 1 & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

Eight different basic feasible solutions to this problem were discovered, namely:

$$B_1 = \begin{bmatrix} 0 & 4 & 0 \\ 0 & 1 & 0 \\ 0 & -3.5 & 0 \end{bmatrix} \qquad B_2 = \begin{bmatrix} 4.308 & 1.846 & 0 \\ 0 & 1 & 0 \\ -7.000 & 0 & 0 \end{bmatrix}$$

$$B_3 = \begin{bmatrix} 5.714 & 1.143 & 0 \\ 0 & 1 & 0 \\ -9.286 & 1.143 & 0 \end{bmatrix} \qquad B_4 = \begin{bmatrix} 0 & 1.143 & 5.714 \\ 0 & 1 & 0 \\ -15.000 & 1.143 & 5.714 \end{bmatrix}$$

$$B_5 = \begin{bmatrix} 0 & 1.455 & 5.091 \\ 0 & 1 & 0 \\ -7.000 & 0 & 0 \end{bmatrix} \qquad B_6 = \begin{bmatrix} 0 & 0.842 & 6.316 \\ 0 & 1 & 0 \\ -8.684 & 0.842 & 0 \end{bmatrix}$$

$$B_7 = \begin{bmatrix} 0 & 1.846 & 4.308 \\ 0 & 1 & 0 \\ -11.308 & 0 & 4.308 \end{bmatrix} \quad B_8 = \begin{bmatrix} 0 & 0.429 & 7.143 \\ 0 & 1 & 0 \\ 0 & 0.429 & -7.857 \end{bmatrix}$$

Convex combinations of these matrices can be used to produce countless realizable admittance matrices yielding networks with the proper transfer resistance. A few of these follow:

$$B_9 = 0.4B_3 + 0.6BB_4 = \begin{bmatrix} 2.286 & 1.143 & 3.429 \\ 0 & 1 & 0 \\ -12.714 & 1.143 & 3.429 \end{bmatrix}$$

$$YB_9 = \begin{bmatrix} 34.000 & -5.000 & -24.000 \\ -5.000 & 5.000 & 0 \\ -24.000 & 0 & 24.000 \end{bmatrix} = \hat{Y}_1$$

$$B_{10} = 0.5B_1 + 0.5B_4 = \begin{bmatrix} 0 & 2.571 & 2.857 \\ 0 & 1 & 0 \\ -7.500 & -1.179 & 2.857 \end{bmatrix}$$

$$YB_{10} = \begin{bmatrix} 25.000 & -2.500 & -20.000 \\ -2.500 & 3.750 & 0 \\ -20.000 & 0 & 20.000 \end{bmatrix} = \hat{Y}_2$$

$$B_{11} = 0.4B_6 + 0.6B_3 = \begin{bmatrix} 3.429 & 1.023 & 2.526 \\ 0 & 1 & 0 \\ -9.045 & 1.023 & 0 \end{bmatrix}$$

$$YB_{11} = \begin{bmatrix} 17.579 & -4.158 & -9.263 \\ -4.158 & 5.000 & -0.842 \\ -9.263 & -0.842 & 10.947 \end{bmatrix} = \hat{Y}_3$$

$$B_{12} = 0.9B_8 + 0.1B_4 = \begin{bmatrix} 0 & 0.500 & 7.00 \\ 0 & 1 & 0 \\ -1.500 & 0.500 & -6.500 \end{bmatrix}$$

$$YB_{12} = \begin{bmatrix} 5.000 & -0.500 & -4.000 \\ -0.500 & 5.000 & -4.500 \\ -4.00 & -4.500 & 13.000 \end{bmatrix} = \hat{Y}_4$$

The realizations of these appear in Figs. 2 through 5.

RLC CIRCUITS

Although the method has been given for resistive networks it is a simple matter to extend it to RLC circuits. What needs to be done is to find transformation coefficients b_{ij} which simultaneously satisfy conditions of the form of (3) for a transformed conductance matrix G, for a transformed capacitance matrix C, and for a transformed reciprocal inductance matrix Γ. For a given number of b's it may not be possible to find values that satisfy all the constraints simultaneously. One alternative that can be tried is to augment the size of the Y matrix and transform the larger matrix as indicated in Ref.2. For one-element-kind networks the augmentation is beneficial since for large n the number of b's grows approximately as n^2

while the number of constraints grows as $n^2/2$ (only the equality constraints are counted since the inequality constraints create slack or surplus variables.) For two-element-kind networks the b's grow approximately at the same rate as the constraints while for three-element-kind networks the number of constraints grow faster than the number of b's. The question for two and three-element kind networks is not definite because what really matters is not the number of constraints but rather the number of <u>independent</u> constraints. This is a function of the particular numbers in each problem and of how the matrices are augmented. For networks which are not fully connected with an R an L and a C between every pair of nodes the zeros play an important role.

IMPLEMENTATION

It is relatively straightforward to implement a digital computer program that will automatically generate RLC equivalent circuits using the methods given in this paper. One may start with a nodal admittance matrix of the form

$$Y = G + sC + s^{-1}\Gamma \qquad (6)$$

where G, C, Γ are constant conductance, capacitance, and reciprocal inductance matrices and s is the complex frequency. The matrix Y in the form given in (6) can be easily obtained by giving a program a coded description of the circuit. The equality and inequality conditions for each of the three matrices of (6) can be expressed in matrix form as shown in (5). It is, of course, very desirable that the matrix be automatically generated by the program since it would be quite laborious to obtain it by hand. In order to produce a general program, one may start by assuming all the b's to be free. The particular form of the matrix B for a given objective is taken care of by adding equality conditions to certain elements (that is additional conditions making some elements one or zero are written independently). This has the effect of adding rows to the matrix of (5). In this way a single program may be used to generate equivalent circuits having different circuit functions invariant.

The algorithms for generating the matrix of (5) are relatively simple. For example for an N-node circuit, if the matrix of (5) is stored in an array P(I,J), the generation of the G portion of the matrix corresponding to the equality constraints to make the final matrix symmetric could be done by the following "quasiFORTRAN" program:

```
SET ALL P(.,.) TO ZERO
DO 10 I=2,N
II=I-1
DO 10 K=1,II
DO 5 J=1,N
L1=(K-1)N+J
L2=(I-1)N+J
P(JJ,L1)=G(I,J)
P(JJ,L2)=-G(K,J)
JJ=JJ+1
IF(JJ≥(N^2-N)/2) EXIT LOOP
5   CONTINUE
10  CONTINUE
```

Once the inequality-equality matrix and right side vector of (5) have been generated, any linear programming routine may be used to obtain a feasible solution. If the program is not very ambitious it may stop when it gets one feasible solution. One problem that may arise is that the feasible solution results in a singular B matrix. Since $\hat{Y}=YB$, this implies a singular $\hat{Y}$ which may give trouble. The trouble is due to the fact that the new impedance matrix $\hat{Z}$ corresponding to the inverse of $\hat{Y}$ does not exist and the transformed network is degenerate. Depending whether the trouble appears at the ports of interest or not, the new network may not be equivalent to the original one. To avoid this problem a more sophisticated program obtains several basic feasible solutions (if they exist) and interpolates between them as indicated above. Even if the basic feasible solutions correspond to singular B matrices, their linear combinations generally will not be singular. To obtain more than one basic feasible solution there are several ways of proceeding. A simple one when a packaged linear programming routine is available and it is not easy to modify it is to resolve the problem several times with different cost coefficients. Although the optimization portion of the problem is not of interest, any optimal solution will also be feasible, hence by changing the cost coefficients we have hope of getting different feasible solutions. The trouble is that 1) we have to resolve the problem completely for each new feasible solution; and 2) we have no guarantee that different cost coefficients will not yield a feasible solution previously obtained.

A more efficient approach is to get into the program and examine the Simplex Tableau at the point at which the first basic feasible solution is obtained. All those variables for which $z_j-c_j = 0$ can be put in the basis without leaving the feasible space. Also all the variables in the basis whose value is zero can be interchanged with any other regular variable not in the basis. By programming these two conditions additional basic feasible solutions may be obtained in one step. However, unless one is careful, one can easily start repeating solutions obtained previously. It is therefore necessary to keep track of the ones that have been already obtained to avoid their recalculation. A method proposed by Charnes[4] based on the maze solving algorithm can find all the basic feasible solutions repeating at most twice each solution.

Once several or all basic feasible solutions are available an infinity of B matrices may be generated by interpolation. (Observe in the illustrative example given above that several of the B's are singular, but that the various linear combinations gave interpolated B's which are not singular.)

Finally the equivalent circuits can be generated by multiplying the original Y matrix by B to obtain $\hat{Y}$ and synthesizing the $\hat{Y}$. The output of the program may be a coded description of the equivalent circuit. A program such as the one described is presently being implemented at the Universidad Nacional Autónoma de México.

For moderate size circuits the CPU time involved in generating a set of equivalent circuits on a computer of the size and speed of the Burroughs 5500 is of the order of one second. This allows the program to be run conveniently in a time-shared

environment in which the terminal times may be the CPU times multiplied by a factor of the order of 50.

CONCLUSIONS

The paper has presented a practical computer method of generating realizable transformerless equivalent circuits. The starting point is an arbitrary matrix having the desired impedance characteristics at the ports of interest. The matrix may or may not be realizable. The method is relatively easily implementable in a computer program that accepts a circuit described in code and outputs equivalent circuits also described in code. The user does not have to deduce any equations or write special programs for each case.

If more than one solution exists the method yields a continuum of equivalent networks. These additional networks are obtained very simply and efficiently.

Possible applications of the method include among others the synthesis of resistor, RC, and RLC N-ports all three important problems still defying complete solution. By forcing certain entries in the transformed matrix to be zero or to be equal to others, several practical constraints may be imposed on the topology or physical symmetry of the final circuits. This is important in the less academic but important practical problem of the synthesis of DC transistor-resistor N-ports with more than N+1 terminals[5]. These and other applications of the approach presented in the paper will be considered in detail elsewhere.

REFERENCES

1. Cauer, W., "Vierpole", Eleck. Nachr. Tech., Vol 6, pp. 272-282, 1929.

2. Murray-Lasso, M. A., "A General Transformation with Application to Circuit Theory", Journal of the Franklin Institute, Vol. 285, No. 2, pp. 79-97, Feb. 1968.

3. Dantzig, G. B., Linear Programming and Extensions, Princeton University Press, Princeton, N. J., 1963.

4. Charnes, A., Cooper, W. W., and A. Henderson, An Introduction to Linear Programming, Wiley, New York, 1953.

5. Murray-Lasso, M.A., and E. B. Kozemchak, "On the Computer Synthesis of Linear Grounded Transistor-Resistor N-Ports with Prescribed Global Sensitivities", Proceedings Second Hawaii International Conference on System Sciences, Honolulu, Hawaii, pp. 575-578, Jan. 1969.

ACKNOWLEDGEMENT

The research reported in this paper was done under contract for the National Aeronautics and Space Administration, Electronics Research Center (NASA/ERC)

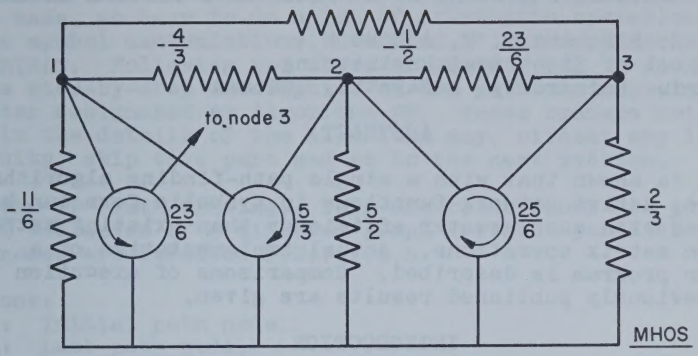


Fig 1 Realization of Y of Eq. (4)

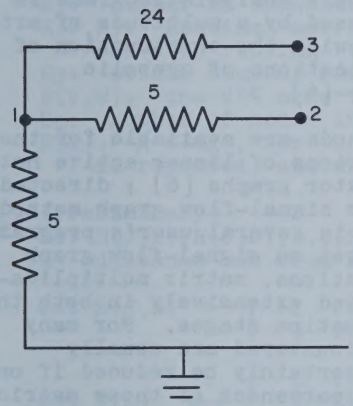


Fig 2 Realization of $\hat{Y}_1$

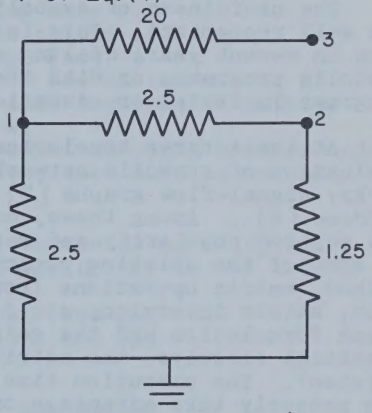


Fig 3 Realization of $\hat{Y}_2$

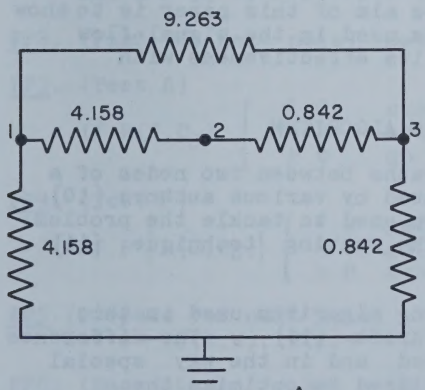


Fig 4 Realization of $\hat{Y}_3$

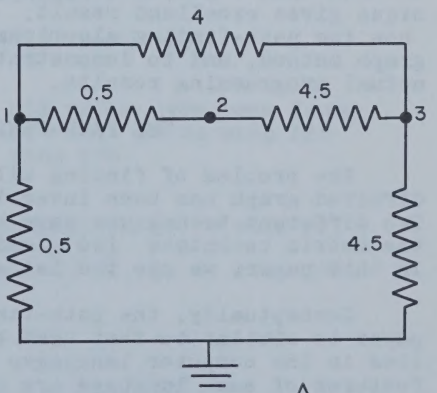


Fig 5 Realization of $\hat{Y}_4$

SYMBOLIC NETWORK FUNCTIONS BY A SINGLE PATH-FINDING ALGORITHM

P. M. LIN and G. E. ALDERSON
School of Electrical Engineering
Purdue University, Lafayette, Indiana

ABSTRACT

It is shown that with a single path-finding algorithm as the core, active network functions in symbolic form can be generated with much greater efficiency than existing methods based on matrix operations. Actual implementation of a computer program is described. Comparisons of execution time with previously published results are given.

INTRODUCTION

The usefulness of symbolic network analysis programs is now well recognized. This is evidenced by a multitude of articles in recent years dealing either with the construction of symbolic programs, or with the applications of symbolic programs in design or education [1]—[6] .

At least three topological methods are available for the evaluation of symbolic network functions of linear active networks: signal-flow graphs [7] ; unistor graphs [8] ; directed k-trees [9] . Among these, only the signal-flow graph method has enjoyed popularity and resulted in several user's programs. In most of the existing programs based on signal-flow graph method, matrix operations (row operations, matrix multiplication, matrix inversion, etc.) are used extensively in both the graph formulation and the gain evaluation stages. For many practical circuits, the matrices encountered are usually 'sparse'. The execution time will certainly be reduced if one can properly take advantage of the sparseness of those matrices.

In exploring various ways of reducing the computing time, we find that a path-finding algorithm based on routing technique gives excellent result. The aim of this paper is to show how the path-finding algorithm is used in the signal-flow graph method, and to demonstrate its effectiveness with actual programming results.

THE PATH-FINDING ALGORITHM

The problem of finding all paths between two nodes of a directed graph has been investigated by various authors [10] . Two different techniques have been used to tackle the problem: the matrix technique [10] and the routing technique [11] . In this paper, we use the latter.

Conceptually, the path-finding algorithm used in this paper is similar to that used by Kroft [11] . The difference lies in the computer languages used, and in the way special features of each language are utilized to optimize the

This work was supported by the National Aeronautics and Space Administration, Grant NGL 15-005-021.

algorithm. Kroft used SNOBOL, a string manipulation language. In our case, we have to do a lot of arithmetic operations besides symbol manipulations. We find it more appropriate to use FORTRAN. Following the style of Knuth [12], we give below a step-by-step description of the path-finding algorithm, hereafter designated as Algorithm PF. Those readers not interested in the details of the algorithm may, without any loss of continuity, skip this part and go to the next section.

Algorithm PF (Path-finding): This algorithm finds all paths between two nodes of a directed graph (without parallel edges) whose nodes are lebelled 1,2,.....N.

Notations:

- I: Initial path node.
- L: Last path node.
- N: Number of nodes in graph.
- E_J: Number of branches leaving node J.
- R(J,M): Routing matrix, also called one-step routing table in [10].
- C_J: Column counter for the J^{th} row of the matrix R.
- P(V,W): The V^{th} node in the node sequence of path W.
- U_W: Number of nodes in path W.
- F(K): A flag, for the purpose of checking whether node K is repeated, and whether the last node is reached.

PF1. (Preliminary)
Set R(J,1),R(J,2),...,R(J,E_J) to the group of E_J nodes of distance one from node J, one by one in any order.

Set $R(J,M) \leftarrow \begin{cases} -1 & \text{for } M=E_J+1 \text{ and } J=I \\ 0 & \text{for } M=E_J+1 \text{ and } J\neq I \end{cases}$

Set $F(K) \leftarrow \begin{cases} 1 & \text{for } K=I \\ 0 & \text{for } K=J \text{ and } J\neq I,L \\ -1 & \text{for } K=L \end{cases}$

Set $C_J \leftarrow 1$ for J=1,2,...,N

Set $W \leftarrow 1$, $V \leftarrow 2$, $J \leftarrow I$, $P(1,1) \leftarrow I$

PF2. (Find the next node) Set $P(V,W) \leftarrow R(J,C_J)$

PF3. (Test R)
If $R(J,C_J)$ $\begin{cases} < 0 & \text{stop; all paths have been found} \\ = 0 & \text{set } F(J) \leftarrow 0, \text{ go to step PF6} \\ > 0 & \text{go to step PF4} \end{cases}$

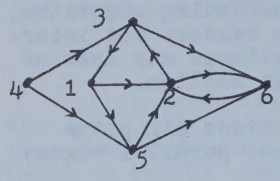
PF4. (Test F)
If $F\{R(J,C_J)\}$ $\begin{cases} < 0 & \text{path completed; go to step PF7} \\ = 0 & \text{go to step PF5} \\ > 0 & \text{set } C_J \leftarrow C_J+1, \text{ go to step PF2} \end{cases}$

PF5 (Prepare for next node)
Set $J \leftarrow P(V,W)$, $F(J) \leftarrow 1$, $V \leftarrow V+1$, go to step PF2

PF6. (Back step) Set $C_J \leftarrow 1$, $J \leftarrow P(V-2,W)$
$C_J \leftarrow C_J+1$, $V \leftarrow V-1$, go to step PF2

PF7. (Finish path) Set $C_J \leftarrow C_J+1$, $P(K,W+1) \leftarrow P(K,W)$,K=1,..,$U_W-1$, $W \leftarrow W+1$, go to step PF2.

As an example, let us use Algorithm PF to find all paths from node 4 to node 6 in Fig. 1. The routing matrix R upon finishing step PF1, and the node-path matrix P(V,W) after enumerating all paths will be as shown.



$$R=\begin{bmatrix} 5 & 2 & 0 \\ 6 & 3 & 0 \\ 6 & 1 & 0 \\ 5 & 3 & -1 \\ 6 & 2 & 0 \\ 2 & 0 & \end{bmatrix} \qquad P=\begin{bmatrix} 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 5 & 3 & 3 & 3 & 3 \\ 6 & 2 & 2 & 6 & 1 & 1 & 1 \\ & 6 & 3 & & 5 & 5 & 2 \\ & & 6 & & 6 & 2 & 6 \\ & & & & & 6 & \end{bmatrix}$$

Fig. 1

THE USE OF ALGORITHM PF IN FLOW-GRAPH CIRCUIT ANALYSIS

The major steps required in developing a signal-flow graph circuit analysis program are clearly explained in [13] . For the lack of space, we shall omit all introductory material, and describe only those features which are unique to the path-finding approach. The network model is to consist of R, L, C, independent sources, and linear controlled sources. All four types of controlled sources are admissible. For the sake of brevity, "signal-flow graph" will be abbreviated as SFG.

Finding a tree

The formulation of a primitive SFG starts with the choice of a tree. This is done with Algorith PF as follow: Starting with a null graph, the tree is grown by adding branches successively. Independent voltage sources and controlled voltage sources are the first ones to be added. Then come the passive RLC elements, in any order. In choosing the $(J+1)^{th}$ branch, we test whether a path exists in the graph formed by the previous J branches of the tree, between the two terminal nodes of $(J+1)^{th}$ branch. If so, the branch under consideration is disqualified. If not, the $(J+1)^{th}$ branch is added to the tree. Let n be the number of nodes of the network graph. When n-1 branches are successfully chosen by the above process, we have obtained a tree. Note that independent and controlled current sources are always excluded from the tree. Also, the problem of 'optimum tree' [14] has not been considered in the above procedure.

Formulating the Compact SFG

In the primitive SFG used by many authors, each network branch calls for two nodes in the SFG-- one for the branch voltage and one for the branch current. In using Algorithm PF, the computing time increases significantly with the number of nodes of the graph, even though some of them are redundant by inspection. Therefore, it is desirable to make the number of nodes in the SFG as small as possible. This is achieved through the use of a 'compact SFG' whose node variables consists of tree branch voltages and link currents only, except for the situations discussed in (2) and (3) below.

The compact SFG is constructed with the aid of the path-finding algorithm according to the following rules:

(1) For each link k, the PF algorithm yields a path through the tree and defines a fundamental circuit C_k. Let b_i, i=1,2,...,m be the tree branches in C_k. Two sets of branches in the compact SFG will be created.

(a) For each passive branch in the set b_i, i=1,2,...,m, a directed branch in the SFG is formed from node I_k to node V_{bi}, with weight equal to the impedance of branch b_i, prefixed with a proper sign (positive, if the directions of k and b_i concur in C_k; negative otherwise).

(b) If the link k is a passive branch, a directed branch in the SFG is formed from each node V_{bi}, i=1,2...,m, to node I_k , having weight equal to the admittance of link k, prefixed with a proper sign (negative, if the directions of k and bi concur in C_k; positive otherwise).

(2) If any of the four types of controlled sources are present, a directed branch is created in the SFG from the controlling variable to the controlled source, having weight equal to the constant of proportionality (g_m, beta, etc.). If the controlling variable is a link voltage or a tree branch current, one more node is added to the SFG to represent this controlling variable x. x is then expressed in terms of tree branch voltages and link currents either through a simple immittance relationship or through the use of PF algorithm.

(3) If the desired output y is neither a tree branch voltage nor a link current, then one node is added to the SFG to represent y. y is then expressed in terms of tree branch voltages and link currents either through a simple immittance relaship or through the use of PF algorithm.

It should be pointed out Rule (1) is in fact an application of the relationship between fudamental circuit matrices and fudamental cutset matrices.

Finally, the compact SFG is "closed" by adding a branch with a symbolic weight, directed from the output to the input node. See [1] , for example, for a discussion of the advantage of using closed SFG.

As an example, consider the network shown in Fig. 2(a), If the tree is chosen to consist of branches 1,2,3,and 4, , the corresponding compact SFG is shown in Fig. 2(b).

Generating First Order Loops

Let the nodes of the SFG be labelled 1,2,...,N. All first order loops which contain node J (J=1 initially) are found by applying the PF algorithm directly specifying node J both as the initial node and the last node . All branches going into node J are then removed and the process repeated for node J+1.

For the SFG of Fig. 2(b), the above procedure generates seven first order loops in the following sequence:

(15263741),graph modification, (252), (262), graph modification, (363), (373), graph modification, (474),(484), end.

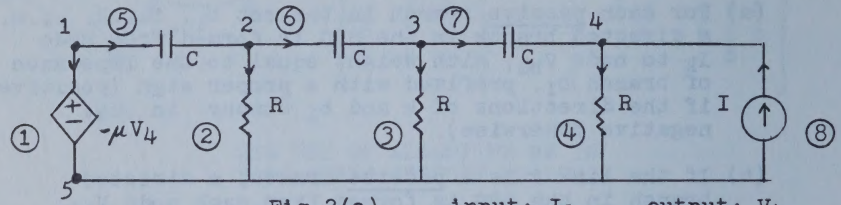


Fig 2(a) input: I_8 output: V_4

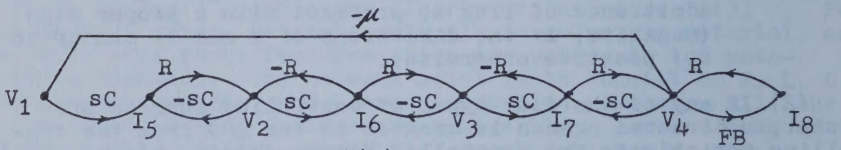


Fig. 2(b)

Generating Second Order Loops

Let the first order loops be numbered 1,2,...,L. For each loop K, all higher numbered loops not touching loop K are found by comparing codes (see the next section). The results are stored in a 'table of nontouching pairs'. The format of this table is illustrated below for the SFG of Fig. 2(b).

Loop K	Higher numbered loops not touching loop K
1	none
2	4,5,6,7
3	5,6,7
4	6,7
5	7
6	none

Generating higher order loops

The problem is reduced to a group of path-finding problems. To find all orders of loops containing loop J, but not containing any lower numbered loops, we construct a directed graph G_L as follows:

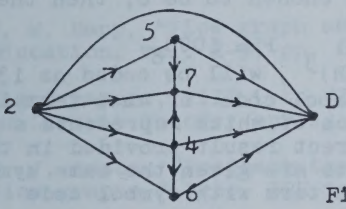
(1) G_L has one node labelled D, and the remaining nodes numbered in accordance with the entries of the J^{th} row of the 'table of nontouching pairs'.

(2) One directed branch is created from each numbered node of G_L to node D.

(3) A directed branch is created from each node K of G_L to each higher numbered node M (M > K) if M is in the K^{th} row of the table of nontouching pairs.

Having constructed G_L, we use Algorithm PF to find all paths from node J to node D. The list of all paths, with the symbol D disregarded is identical to the list of all orders of loops containing loop J, but not containing lower numbered loops. When the procedure is repeated for J=1,2,...,L-1, we obtain all orders of loops in the SFG, among which the first and second order loops have been generated previously.

As an example, consider the SFG of Fig. 2(b). For J=2, the corresponding G_L and the list of all paths from node 2 to to node D are shown in Fig. 3. One can verify easily that the list gives all orders of loops containing loop 2 but not loop 1.



2D, 24D, 246D, 247D, 25D
257D, 26D, 27D

Fig. 3

CODING SCHEMES

Numerical codes are used to determine second order loops and to manipulate symbolic weights.

Second Order Loops (nontouching pairs)

As each first order loop is generated, it is assigned an integer code whose binary representation shows the set of nodes in the loop. For example, if loop 3 contains nodes (1,4,5), and loop 4 contains nodes (2,3,5), then the codes are

$N3 = (000000011001)_2 = 25$
$N4 = (000000010110)_2 = 22$

To determine whether two loops touch or not, the masking operator .AND. is used. For the example above,

$N3 \text{ .AND. } N4 = (000000010000)_2$

The result is not zero, indicating that loops 3 and 4 touch.

Symbolic Weights of Loops

Each branch weight in the SFG is of the form

Constant · Symbol · s^n

Thus, three parameters are ascribed to a branch to define its weight- one each for the constant, symbol, and power of s. Distinct symbols in the branch weights are coded as follows:

no symbol	0	base B
symbol #1	10	
synbol #3	100	
etc.		

where the base is any number chosen from $2,4,...,2^m$. (The

advantage of using a base $B > 2$ will be seen shortly). Using this technique, a loop of any order which contains branches b_i, $i = 1,2,\ldots$, can be completely described by a constant which is the product of the individual constants, a symbol code which is the sum of the individual codes, and a power of s corresponding to the sum of each exponent.

Each loop (of any order) contributes to a term in the network function. As each loop (of any order) is generated and coded, it is compared with existing terms. If a term with the same symbol code and power of s exists, then the constant of the term is updated by adding to it the constant of the new loop. (This is an important step toward reducing the storage requirement.) Otherwise, a new term is created.

Again, consider the SFG of Fig. 2(b). There are 4 symbols R, C, μ, and FB. If the base B is chosen to be 8, then these 4 symbols may be coded as follows:

$R = 1_8$, $C = 10_8$, $\mu = 100_8$, $FB = 1000_8$

Loop 1, which has a weight $-\mu(sCR)^3$, will be coded as 133_8. In the final output stage, the symbol code of each term is decomposed into its components each of which represents a symbol. The decomposition gives correct result provided in the network no more than (B-1) elements are given the same symbol. For the example discussed above, a term with symbol code 22_8 will be decoded as $(RC)^2$.

SOME RESULTS FROM SNAP

A computer program called SNAP (Symbolic Network Analysis Program) which embodies all the concepts discussed in this paper has been written. Running on CDC6500, the program requires approximately 20K storage locations for most networks of up to 35 branches. Although the program finds a tree automatically to formulate the compact SFG, the user can choose a particular tree by simply having the passive branches of that tree read in first. Two examples are included here to illustrate the input , output formats, and the execution time required. The first example, shown in Fig. 4, has been discussed by several authors. A decade ago, researchers took a rather dim view of symbolic network programs. To handle a problem such as Fig. 4, the time was 4 hours (estimate) in [15] ; 4 minutes (estimate) in [16] ; 7.49 seconds (for listing trees only) in [17] ; and in the present paper, 3.01 seconds for obtaining the complete answer as a rational function of s. The network of Fig. 5 was used as an example in [2] . The execution time was 2.37 minutes (with compilation), with all elements taking on numerical values. With SNAP, the execution time is 0.71 second, with 3 elements kept in symbols.

The improvement in execution time is thought mainly due to the algorithm used in the program. We have avoided matrix operations, but used extensive searching operations. In [3], we see that part of NASAP is modified along this line. There is strong indication that routing technique is superior for symbolic programs than matrix technique.

REFERENCES

1. W. W. Happ and R. M. Carpenter, "Algorithms for the dichotomus representation of macrocircuits," Proc. 4th Allerton Conf. on Circuit and System Theory, Univ. of Ill., 1966.

2. F. M. Verduyn and S. P. Chan, "Analysis of active networks by digital computer using a flowgraph technique," Proc. 2nd Asilomar Conf. on Circuit Theory, November, 1968.

3. R. E. Bach, Jr. and R. S. Schwartz, " A parameter modification algorithm for NASAP," Proc. 12th Midwest Symp. on Circuit Theory, April, 1969.

4. R. E. Bach, Jr., G. F. Currin and P. A. Olendaenki, "A circuit analysis program for educational use," Proc. 2nd Asilomar Conf. on Circuit Theory, November, 1968.

5. W. W. Happ, "Flow graph as a teaching aid," IEEE Trans. on Education, vol. E-9, pp. 60-79, June, 1966.

6. Eight papers, VIII.1 to VIII.8, in Proc. 12th Midwest Symp. on Circuit Theory, April, 1969.

7. S. J. Mason, "Feedback theory-further properties of signal flow graphs," Proc. IRE,vol. 44, pp. 920-926, July, 1956.

8. S. J. Mason, "Topological analysis of linear nonreciprocal networks," Proc. IRE,vol. 45, pp. 829-838, June, 1957.

9. Wai-kai Chen,"Topological analysis for active networks," IEEE Trans. on Circuit Theory, pp. 85-91, March, 1965.

10. I. M. Paz, "Pathfinding through a communication network," Proc. IEE, pp. 76-78, Jan., 1967. And its references.

11. D. Kroft, "All paths through a maze," Proc. IEEE, vol. 55, pp. 88-90, January, 1967.

12. D. E. Knuth, The Art of Computer Programming, vol. 1, Fundamental Algorithms, Addison-Wesley, 1968.

13. G. W. Zobrist,Ch. 5 in Computer-aided Integrated Circuit Design, edited by G. J. Herskowitz, McGraw-Hill, 1968.

14. G. W. Zobrist and J.E. Barbay,"Distinguishing characteristics of the optimum tree," Proc. 5th Allerton Conf., 1967.

15. J. MacWilliams,"Topological network anslysis as a computer program," IRE Trans. on Circuit Theory, pp. 228-229, September, 1958.

16. E. W. Hobbs, "Topological network analysis as a computer program," IRE Trans. on Circuit Theory, pp. 135-136, March, 1959.

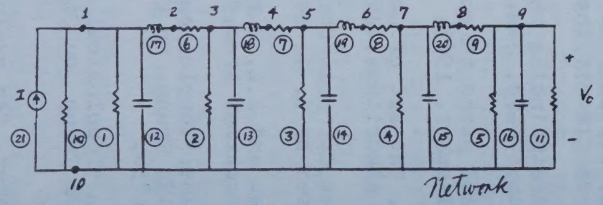
17. E. B. Fernandez, "An evaluation of tree generation methods" Proc. 12th Midwest Symp. on Circuit Theory, April, 1969.

```
C     10=NODE 20=BRANCH RLC LADDER NETWORK
      10    21
      21     5
R     1      1     10   R1=1.
R     2      3     10   R2=1.
R     3      5     10   R3=1.
R     4      7     10   R4=1.
R     5      9     10   R5=1.
R     6      2      3   R6=1.
R     7      4      5   R7=1.
R     8      6      7   R8=1.
R     9      8      9   R9=1.
R     10     1     10   R10=1.
R     11     9     10   R11=1.
C     12     1     10   C12=1.
C     13     3     10   C13=1.
C     14     5     10   C14=1.
C     15     7     10   C15=1.
C     16     9     10   C16=1.
L     17     1      2   L17=1.
L     18     3      4   L18=1.
L     19     5      6   L19=1.
L     20     7      8   L20=1.
I     21    10      1   I1
```

10 nodes, 21 branches
input (21), output (5)

(5) is of type R, connected from 9 to 10
R5 = 1 ohm

Data cards, format illustrated above.



Answer

$$V_o/I = V_5/I_{21}$$

$$= 1/(144 + 566s + 1051s^2 + 1211s^3 + 956s^4 + 538s^5 + 217s^6 + 61s^7 + 11s^8 + s^9)$$

Fig. 4

NUMERATOR POLYNOMIAL

```
COLUMN          SYMBOL FOR GIVEN COLUMN
  1                 1          1 / 1

EXPONENT                CONSTANT COEFS. IN THE POLYNOMIAL
                          1
  0           1.00000E+00
  1           0.
  2           0.
  3           0.
  4           0.
  5           0.
  6           0.
  7           0.
  8           0.
  9           0.
```

****

DENOMINATOR POLYNOMIAL

```
COLUMN          SYMBOL FOR GIVEN COLUMN
  1                 0          1 / 1

ADD S**( 4) TO DENOMINATOR

EXPONENT                CONSTANT COEFS. IN THE POLYNOMIAL
                          1
  0           1.44000E+02
  1           1.21100E+03
  2           5.38000E+02
  3           9.55000E+02
  4           1.05100E+03
  5           7.17000E+02
  6           5.66000E+02
  7           6.10000E+01
  8           1.00000E+01
  9           1.00000E+00
EXECUTION TIME IN SECONDS.          8.01000000
```

```
C  2 STAGE AMPLIFIER
        4    13
        1     5
E       1     1     6    E1
C       2     1     2    C2=1.
R       3     2     3    R3
R       4     4     6    R4=1.
R       5     5     6    R5=1.
G       6     1     6    G6=1.
C       7     1     6    C7=1.
VC      8     2     6    GM                    4
C       9     3     4    C9=1.
G      10     3     4    G10=1.
CC     11     4     5    A                     3
C      12     4     5    C12=1.
G      13     4     5    G13=1.
```

Data cards
6 nodes, 13 branches
input ①, output ⑤

R_3, GM, A kept in symbolic
other element values = 1.

controlled source

controlled source

NUMERATOR POLYNOMIAL

COLUMN		SYMBOL FOR GIVEN COLUMN
1	2	1 / 1
2	23	A R3 / R3
3	6	R3 / 1
4	10	GM / 1
5	31	A GM R3 / R3
6	14	GM R3 / 1

EXPONENT	CONSTANT COEFS. IN THE POLYNOMIAL					
	1	2	3	4	5	6
1	1.00000E+00	1.00000E+00	0.	-2.00000E+00	-1.00000E+00	0.
2	2.00000E+00	1.00000E+00	0.	-1.00000E+00	0.	0.
3	1.00000E+00	0.	0.	0.	0.	0.
0	0.	0.	0.	-1.00000E+00	-1.00000E+00	0.

DENOMINATOR POLYNOMIAL

COLUMN		SYMBOL FOR GIVEN COLUMN
1	0	1 / 1
2	21	A R3 / R3
3	4	R3 / 1

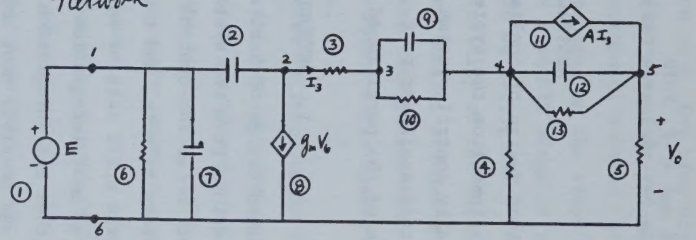
ADD S** (1) TO DENOMINATOR

EXPONENT	CONSTANT COEFS. IN THE POLYNOMIAL		
	1	2	3
1	9.00000E+00	-1.00000E+00	3.00000E+00
2	7.00000E+00	-1.00000E+00	5.00000E+00
3	1.00000E+00	0.	2.00000E+00
0	3.00000E+00	0.	0.

EXECUTION TIME IN SECONDS, .71000000

Fig. 5

Network



Answer (Note that ⑥ and ⑦ are redundant)

$$\frac{V_0}{E} = \frac{V_5}{V_1} = \frac{s^3 + (2+A-g_m)s^2 + (1+A-2g_m-Ag_m)s - g_m - Ag_m}{(1+2R_3)s^3 + (7-A+5R_3)s^2 + (10-A+3R_3)s + 3}$$

FAST COMPUTER ANALYSIS OF NONLINEAR SYSTEMS

R. H. ALLEN and A. F. FATH
Boeing Scientific Research Laboratories, Seattle, Washington

ABSTRACT

Computer aided design of nonlinear differential systems require many response calculations from the system model $\dot{x} = F(x,u)$. Many nonlinear systems, especially control systems, can be separated into a stiff linear part and a nonstiff nonlinear part. Fast numerical integration techniques taking advantage of the partitioned form are presented.

I. INTRODUCTION

Computer aided design of nonlinear differential systems requires many response calculations from the system model $\dot{x} = F(x,u)$ where x is the state vector and u is the control vector. Accurate models include both short and long term effects, and thus have widely separated or stiff eigenvalues. For large stiff sets of equations the analysis takes excessive execution time as most integration techniques (including the Runge Kutta rules) require that the integration time step be limited by the magnitude of the largest and possibly most uninteresting eigenvalues.

Many nonlinear differential systems, especially control systems, can be separated into a linear and a nonlinear part of the form $\dot{x} = Ax + F(x,u)$ where all the stiff eigenvalues are in the linear part represented by the matrix A. This partitioning is possible for most cases because the nonlinear equations usually describe the motion of the system to be controlled (e.g. airplane, missile, hydrofoil, etc.) which has time responses much slower than those of the controlling sensors, compensators and actuators. Further, most sensors, compensators and fast actuators are designed to operate in a linear or almost linear region of their response curves. Nonlinearities, such as physical limits on actuator travel, can most often be incorporated into the plant equation as input saturation. Thus the total system including feedback which is represented by $F(x,u)$ can be partitioned so that the large eigenvalues occur in the linear part.

In the numerical response calculation for nonlinear differential systems, there are three types of errors to consider; truncation error, roundoff error, and error due to numerical instability of the integrating rule. Roundoff error can be controlled by use of sufficient word length and programming techniques such as inner-product accumulation. Truncation

error is the one step error due to the discrete approximation to the integral. This error is a function of the integration rule, the time step size, and the state of the system. For a particular integration rule and step size, bounds for the truncation error can be determined, however, actual truncation error can be significantly less depending on the state of the system. For example, if the state of the system is a stable singular point of the nonlinear equations, truncation error could be negligible. The third type of error, that arising from numerical instability of the integration rule, is often dismissed because truncation error control schemes will normally force the step size small enough so that the integration rule is stable. For stiff systems this may cause an unreasonably large number of integration steps to be required by some rules.

Numerical stability of integration rules [1,2] normally depends on the product of the maximum modulus eigenvalue (of the Jacobian) and the time step size T. For example, for a Runge Kutta fourth order rule $|\lambda \max| T$ must be less than 2.8. Thus even though the truncation error might be negligible for larger step sizes, T must remain small to guarantee the computed solution does not grow without bound.

The purpose of this paper is to present several integration rules which are stable over much larger regions than standard rules and thus allow the step size to be increased when the truncation error allows. For these rules, the region of numerical stability depends on the maximum modulus eigenvalue of the nonlinear part only, which due to partitioning may be orders of magnitude less than that of the composit system.

II. RESPONSE CALCULATION

Transformation

Assume a system of the form $\dot{x} = Ax + f(x,u)$ where stiff eigenvalues are only in the linear part A. Transforming the state variables, $x = e^{AT}y$, gives a new set of nonstiff differential equations, $\dot{y} = G(y,u)$ obtained as follows,

$$\dot{x} = Ax + f(x,u) \tag{1}$$

$$Ae^{AT}y + e^{At}\dot{y} = Ae^{At}y + f(e^{At}y,u) \tag{2}$$

$$\dot{y} = e^{-At}f(e^{At}y,u) \tag{3}$$

The dynamic eigenvalues of a nonlinear system are the eigenvalues of the instantaneous Jacobian. The eigenvalues of the original system are the eigenvalues of $A + f_x(x,u)$. The eigenvalues of the transformed system are the eigenvalues of

$$G_y = e^{-At} f_y = e^{-At} f_x x_y = e^{-At} f_x(x,u) e^{At} \tag{4}$$

Since G_y is a similarity transformation of f_x, the eigenvalues of the transformed system, G, are the eigenvalues of the nonlinear part of the original system, f.

The transformed system of equations can be integrated by fast explicit integration techniques without requiring small time steps to prevent numerical instabilities due to stiff eigenvalues. Integrating $\dot{y}$ over one time step, $t_{n+1} - t_n = T$, gives

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} G(y,u)dt \tag{5}$$

Premultiplying by $e^{At_{n+1}}$ gives the convolution integral sequential difference approximation to x_{n+1}.

$$x_{n+1} = e^{AT} x_n + \int_{t_n}^{t_{n+1}} e^{A(t_{n+1}-t)} f(x,u)dt \tag{6}$$

Numerical stability is maintained by using either an exact analytic expression for e^{AT} or by using the numerically stable Pade approximations. Table 1 presents some symmetric Pade approximations to e^{AT}. Numerical techniques for the analytical determination of e^{AT} using similarity transformations are similar to those presented in [3].

Table 1. Numerically stable symmetric Pade approximations to e^{AT}.

Order	
2	$(I - \frac{1}{2}TA)^{-1} (I + \frac{1}{2}TA)$
4	$(I - \frac{1}{2}TA + \frac{1}{12}T^2A^2)^{-1} (I + \frac{1}{2}TA + \frac{1}{12}T^2A^2)$
6	$(I - \frac{1}{2}TA + \frac{1}{10}T^2A^2 - \frac{1}{120}T^3A^3)^{-1} (I + \frac{1}{2}TA + \frac{1}{10}T^2A^2 + \frac{1}{120}T^3A^3)$
8	$(I - \frac{1}{2}TA + \frac{3}{28}T^2A^2 - \frac{1}{84}T^3A^3 + \frac{1}{1680}T^4A^4)^{-1}(I + \frac{1}{2}TA + \frac{3}{28}T^2A^2 + \frac{1}{84}T^3A^3 + \frac{1}{1680}T^4A^4)$

The above transformed system of equations can be stably integrated by any explicit integration rule such as the Adams-Basforth multistep rules. The Runge Kutta rules are used here because they are self starting and

allow easier error detection and correction. Transformed Runge Kutta rules for order's 1, 2 and 4 are as follows:

$$1 \quad x_{n+1} = e^{AT}[x_n + T\, f(x_n)] \tag{7}$$

$$2 \quad k = Te^{AT} f(x_n) \tag{8}$$

$$x_{n+1} = e^{AT} x_n + \tfrac{1}{2}T\left[e^{AT} f(x_n) + f(e^{AT} x_n + k)\right]$$

$$4 \quad k_1 = T\, f(x_n) \tag{9}$$

$$k_2 = T\, f(e^{\frac{1}{2}AT} x_n + \tfrac{1}{2} e^{\frac{1}{2}AT} k_1)$$

$$k_3 = T\, f(e^{\frac{1}{2}AT} x_n + \tfrac{1}{2} k_2)$$

$$k_4 = T\, f(e^{AT} x_n + e^{\frac{1}{2}AT} k_3)$$

$$x_{n+1} = e^{AT}(x_n + \tfrac{1}{6} k_1) + e^{\frac{1}{2}AT} \tfrac{1}{3}(k_2 + k_3) + \tfrac{1}{6} k_4 .$$

Truncation Error

An example stiff system of equations in a later section of this paper demonstrates how the transformed rules maintain numerical stability for large step sizes even though the truncation errors become large. For a linear system of equations, $\dot{x} = Cx$, the truncation error for a p-th order Runge Kutta rule is

$$e_n = x(t_n) - x_n = \frac{T^{p+1} C^{p+1}}{(p+1)!}\, x_n + O(T^{p+2}) \tag{10}$$

For a transformed Runge Kutta rule applied to $\dot{x} = Ax + f$ if $f = Bx$ then the truncation error for a p-th order rule using a p-th order Pade approximation is dominated by

$$e_n = KT^{p+1} A^{p+1} x_n \tag{11}$$

since $||A|| >> ||B||$. If the approximation to e^{AT} is of order greater than p then

$$e_n = KT^{p+1} A^p B x_n . \tag{12}$$

For T much greater than $||A||^{-1}$ the error gets large so that significant increases in T cannot be made.

For example if A has eigenvalues like -1000 and f_x has eigenvalues like -1 then a standard Runge Kutta rule will require a time step of about .001 to maintain numerical stability. The transformed method will be

numerically stable for any step size up to 1, but the truncation error can be like $T^{p+1}10^{3p}/(p+1)!$. To control this truncation error T might still have to be kept small.

Convolution Integral Partitioning

When the stiff system of equations can be further partitioned so that part of the system has no stiff part as is shown in figure 1 and equation 13, then low error convolution integral approximations can be made.

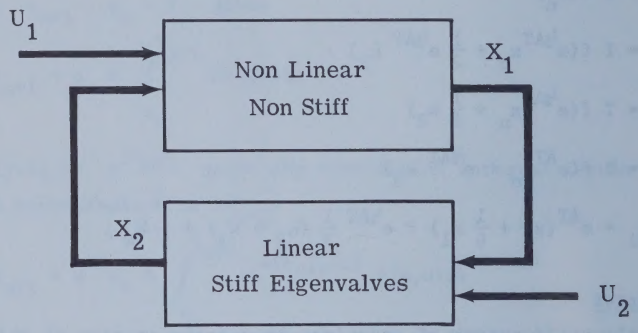


Figure 1. Diagram of Partitioned System

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1(x_1,x_2,u_1) \\ f_2(x_1,x_2,u_2) \end{bmatrix} \tag{13}$$

Numerically stable, low error integration rules can be obtained by applying first and second order explicit one step end point approximations to the convolution integral representation of x_2.

$$x_2^{n+1} = e^{A_2 T} x_2^n + \int_{t_n}^{t_{n+1}} e^{A_2(t_{n+1}-t)} f_2(t)dt \tag{14}$$

This approximation to x_2 is coupled with compatible Runge Kutta rules to approximate x_1.

A first order approximation to the convolution integral can be obtained by expanding f_2 about t_n and then integrating the exponential term exactly.

$$f_2(x(t)) = f_2^n + 0(T) \tag{15}$$

$$\int_{t_n}^{t_{n+1}} e^{A_2(t_{n+1}-t)}dt = A_2^{-1}\,(e^{A_2T} - I) = I_1 \tag{16}$$

The integral I_1 is well defined even if A_2 is not of full rank and can be calculated, as before, using stable Pade approximations or evaluating exact analytical expressions. Using equations (15) and (16), equation (14) becomes

$$x_2^{n+1} = e^{A_2T}x_2^n + I_1 f_2^n + O(T^2). \tag{17}$$

From a first order rule

$$x_1^{n+1} = x_1^n + Tf_1^n \tag{18}$$

Similarly to get a second order rule

$$f_2(x(t)) = f_2^n + \dot{f}_2^n\,[t-t_n] + O(T^2] \tag{19}$$

$$\int_{t_n}^{t_{n+1}} e^{A_2(t_{n+1}-t)}[t-t_n]dt = A_2^{-1}(I_1 - TI) = I_2 \tag{20}$$

A first order approximation to $\dot{f}_2^n$ is needed so evaluating f_2 at x_1^{n+1} and x_2^{n+1} from (17) and (18) gives a first order approximation to f_2^{n+1}.

$$\dot{f}_2^n = \frac{f_2^{n+1} - f_2^n}{T} + O(T^2) \tag{21}$$

Using equations (19), (20), and (21) equation (14) becomes

$$x_2^{n+1} = e^{A_2T}x_2^n + I_1 f_2^n + I_2[f_2^{n+1} - f_2^n]/T + O(T^3) \tag{22}$$

From a second order Runge Kutta rule

$$x_1^{n+1} = x_1^n + \tfrac{1}{2}T(f_1^n + f_1^{n+1}) \tag{23}$$

where again f_1^{n+1} is evaluated at the first order points in equations (17) and (18). Equations (22) and (23) define a partitioned second order rule.

Under the special circumstances when f_2 is a function of only x_1 and the Jacobian of f_2 is readily available (such as when f_2 is linear) then slightly better accuracy can be obtained by using the following approximation to f_2:

$$f_2(x_1(t)) = f_2^n + f_{2x}^n\, f_1^n\,[t-t_n] + O(T^2) \tag{24}$$

Now equation (22) takes the form

$$x_2^{n+1} = e^{A_2 T} x_2^n + I_1 f_2^n + I_2 f_{2x}^n f_1^n \tag{25}$$

Using (18) and (25) f_1^{n+1} can be approximated to be used to define x_1^{n+1} in equation (23). Equations (22) and (23) define a partitioned second order rule when the Jacobian of f_2 is available.

The truncation errors for these partitioned rules are on the order of the truncation errors for a system of equations with eigenvalues on the order of the eigenvalues of the nonlinear parts. For example if A_2 has eigenvalues like -1000 and f_1 and f_2 have eigenvalues like -1 then the truncation error will be like $T^{p+1}.(p+1)!$. To get 3 place accuracy with a first order rule T should be about .04 and with a second order rule T should be about .1.

III. EXAMPLE

In order to illustrate the application of these solution techniques, consider a simple pendulum problem in which the angular rate is measured and used as a signal to a torque motor to provide damping. The nonlinear equations describing the motion of the pendulum are

$$ML\,\ddot{\theta}(t) = -Mg \sin \theta(t) + T(t)/L \tag{26}$$

where M is the mass, L is the length, and T(t) is the torque applied by the torque motor. Let the dynamics of the angle sensor be given by a first order lag with a time constant $1/\tau_s$. Likewise let the torque motor be represented by a first order lag with time constant $1/\tau_m$. To obtain the state equations, let x_1 be θ, x_2 be $\dot{\theta}$, x_3 be the output of the angular rate sensor, and let x_4 be the torque. The equations of motion are then

$$\begin{bmatrix} \dot{x}_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_2 \\ -g/L \sin x_1 - x_4/ML^2 \\ K_s \tau_s x_2 - \tau_s x_3 \\ K_m \tau_m x_3 - \tau_m x_4 \end{bmatrix} \tag{27}$$

For the purpose of this example, let g/L be 1., $1/ML^2 = 0.5$, $\tau_s = 1000.$, $K_s = 1.0$, $\tau_m = 50.$, and $K_m = 3.0$.

Now factor the system into the form

$$\dot{x} = \left[\begin{array}{c|cc} 0 & 0 & \\ \hline 0 & -1000 & 0 \\ & 150 & -50 \end{array}\right] x + \left[\begin{array}{c} x_2 \\ \sin x_1 \quad -.5x_4 \\ \hline 1000\ x_2 \\ 0 \end{array}\right] \tag{28}$$

$$= Ax + F.$$

Note that the Jacobian of F for small x_1 is

$$\frac{\partial F}{\partial x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & -.5 \\ 0 & 1000 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{29}$$

which has eigenvalues of 0.0, 0.0, $\pm j1.0$. The eigenvalues of the composit system for small x_1 are -1000.08, -48.371, $-.77484 \pm j\ .65818$.

For the example system of equations considered here some indications of the acceptability of the suggested integration rules can be made. Figure 2 is a plot of discretization error curves for several rules. The log of the error versus the log of the integration time step has a slope which is equal to the order of the integration rule. Table 2 compares the number of significant figures of accuracy obtained for the integrating rules at several time steps. Also computer execution times are given.

A fourth order Runge Kutta rule had high accuracy for small step sizes, but it went unstable for acceptable step sizes. An order two Runge Kutta rule went unstable for smaller step sizes. The transformed Runge Kutta rules remained stable for large step sizes but the accuracy was poor. The partitioned methods gave acceptable accuracy for reasonable step sizes. Using the Jacobian technique gave about 1 place better accuracy than the time derivative approximation technique. Using a Pade approximation to e^{AT} instead of an exact expression did not degrade the accuracy much, but in some problems this may cause loss of accuracy because large negative Pade exponents approach -1. instead of 0. The partitioned 2nd order rule allows integration to acceptable accuracy about 200 times faster than a standard 4th order Runge Kutta rule for this example.

Table 2. Number of significant figures and timings for example.

Integration rule/time step	10^{-3}	10^{-2}	10^{-1}
Order 4 Runge Kutta	12	unstable	unstable
Order 2 Runge Kutta	unstable	unstable	unstable
Order 2 transformed	3	1	0
Order 2 partitioned	6	4	2
Order 2 partitioned (with Pade approximation)	6	4	2
Order 2 partitioned (with Jacobian)	7	5	3
Execution time in seconds for all order 2 rules (360/44)	30	3	.3

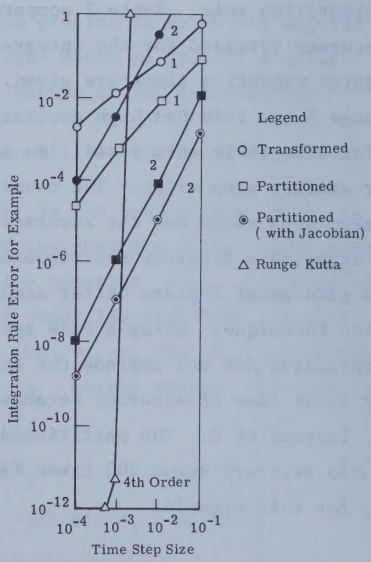


Figure 2. INTEGRATION ERROR CURVES

IV. CONCLUSIONS

In many simulations where one or .1 percent accuracy is sufficient, the larger step sizes allowed by the low order stable integration rules reduces run times appreciably. Control systems for nonlinear plants which use position and rate feedback with compensation usually allow partitioning as indicated in Eq. 13 with f_2 being a function of x_1 and u only. Experience has shown the second order rule to be ideally suited for this problem. Where acceleration feedback is used, f_2 now becomes a function of x_2 in addition to x_1 and u which in some cases reduces the effectiveness of the integration rule.

The rules presented here show advantages over other rules such as standard Runge Kutta only when the truncation errors of the primary variables are small enough to warrant larger step sizes. Where the problem is truncation error limited, the higher order rules offer advantages in error control. It should be noted that in many simulations, however, errors due to numerical instabilities are the limiting factor. For these systems, stable integration rules can significantly reduce computation times.

One obvious example where these rules are particularly effective is in determining steady state operating conditions of nonlinear systems. Here, large errors in the transient solution can be tolerated making stable integration rules with large step sizes quite desirable.

REFERENCES

1. Allen, R. H., Fast Computer Aided Analysis of Nonlinear Electronic Circuits, Proc. Cornell Conference on Computerized Electronics, August 1969.

2. Calahan, D. A., Computer-Aided Network Design (book), McGraw-Hill, New York, 1968, pps. 66-75.

3. Fath, A. F., Computational Aspects of the Linear Optimal Regulator Problem, Preprints of 1969 Joint Automatic Control Conference, University of Colorado, August 1969, pps. 44-49.

COMPUTATIONAL STABILITY IN SYSTEM ANALYSIS

BERTIL S. M. GRANBORG
Department of Electrical Engineering
University of Hawaii
Honolulu, Hawaii 96822

ABSTRACT

The computational stability in the analysis of system dynamics is investigated utilizing nine different integration algorithms. A method is presented to determine the optimum stability of a system of n linear or non-linear differential equations. Some examples verify the theory and numerical calculations are performed on a digital computer.

1. INTRODUCTION

The use of digital computers to simulate transient phenomena in systems and processes has created an increasing interest in digital integration methods. The integration is one of the most important calculations of any transient simulation, and this paper makes a comparison of the numerical stability of different self starting integration algorithms. The basis for these calculations is the z-transform techniques, which have been used by Bekey and Karplus (1), and Gilbert (2), who rely upon the frequence response idea. A more direct method is to form an explicit relationship between the computational time increment and the coefficients of the differential equation system.

2. NUMERICAL STABILITY ANALYSIS

Consider the linear or non-linear n-th order differential equation

$$a_n \frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \text{----} a_1 \frac{dx}{dt} + a_o = f(t)$$

which can be reformulated to yield a system of n linear or non-linear differential equations

$$\frac{dx_1}{dt} = a_{11}x_1 + a_{12}x_2 + \text{----} + a_{1n}x_n + f_1$$

$$\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2 + \text{----} + a_{2n}x_n + f_2$$

$$\vdots$$

$$\frac{dx_n}{dt} = a_{n1}x_1 + a_{n2}x_2 + \text{----} + a_{nn}x_n + f_n$$

Similarly two or more n-th order differential equations can be reformulated to a system of 2n or higher number of first order differential equations.

Any number of such equations is conveniently written in State-Variable form

$\{\dot{x}\} = [A]\{x\} + [f]$, where A is the coefficient matrix.

The numerical solution of these equations utilizes discrete data points implying that the differential equations must be considered as difference equations. Similarly, the integration routines are formulated as difference equations whose z-transform transfer functions are used to investigate the numerical stability of the entire system. See Appendix for further details.

The state space form of the system in z-transforms yield

$\{\dot{x}\} = [B(z, T_D)]\{x\}$, where T_D is the computing increment

The bilinear transformation

$$z = \frac{w + 1}{w - 1}$$

is performed on these equations in order to further study the stability conditions more conveniently in the w-plane. Hence,

$$\{\dot{x}\} = [B_1(w, T_D)]\{x\}$$

Combining this set of equations with the original set of differential equations gives

$$[B_1(w, T_D)]\{x\} = [A]\{x\}$$

$$|B_1(w, T_D)\ [I] - [A]|\{x\} = 0$$

where $[I]$ is the identity matrix.

A non-trivial solution exists if the system determinant is zero;

$$|B_1(w, T_D)\ [I] - [A]| = 0$$

The roots of this equation must be pseudo negative for stability, and this is determined by means of Hurwitz's stability criterion. Relationships between the coefficients of $[A]$ and the time increment T_D can be established. If the system is large, considerable effort to perform the calculations is required, but short cuts are possible. (3).

Considering the polynomial $d_o w^n + d_1 w^{n-1} + \text{----} d_n = 0$,

the system is stable if all coefficients are positive and present, and if each of the Hurwitz's principal minors are greater than zero.

Hurwitz's Determinant is $D_n = \begin{vmatrix} d_1 & d_0 & 0 & 0 & \cdots \\ d_3 & d_2 & d_1 & d_0 & 0 \cdots \\ d_5 & d_4 & d_3 & d_2 & 0 \cdots \\ & & \cdots & \cdots & d_n \\ d_{2n-1} & & & & \end{vmatrix}$

and the principal minors are: $D_1 = \left| d_1 \right|$, $D_2 = \begin{vmatrix} d_1 & d_0 \\ d_3 & d_2 \end{vmatrix}$, $D_3 = \begin{vmatrix} d_1 & d_0 & 0 \\ d_3 & d_2 & d_1 \\ d_5 & d_4 & d_3 \end{vmatrix}$ etc.

By simple algebraic manipulations it can be shown that for n=1, $D_1 > 0$; for n=2, $D_1 > 0$ and $D_2 > 0$; for n=3, $D_2 > 0$; for n=4, $D_3 > 0$; for n=5, $D_2 > 0$ and $D_4 > 0$; etc., are sufficient conditions for the system to be stable if all coefficients in the original equation are positive and present. Hence, it is not necessary to compute all the determinants for a large system.

The method described above to find a relationship between the numerical time increment T_D and the coefficients of the original set of differential equations A is applicable to any system order, if the z-transfer function of the integration algorithm can be formed explicitly.

3. NUMERICAL STABILITY CRITERIA

The integration methods considered in this investigation are: The Rectangular (Forward) (4), the Rectangular (Backward) (4), the Trapezoidal (5), the Heun (6), the Fowler (4,8), the Third Order Runge-Kutta (6), the Fourth Order Runge-Kutta (6), the Runge-Kutta-Gill (5), and the Runge-Kutta-Merson (7).

The numerical stability for a single integrator is found by means of the first order differential equation

$$\dot{x} = \frac{1}{T_1} \left[f(t) - x(t) \right]$$

and the integration algorithm.
For instance the Forward Rectangular Integration formula in difference form is

$$x_n = x_{n-1} + T_D \frac{dx_{n-1}}{dt}$$

where T_D is the integration step size.

The combination of these equations yields the difference equation

$$X_n = \left(1 - \frac{T_D}{T_1}\right) X_{n-1} + \frac{T_D}{T_1} f_{n-1}$$

whose z-transform transfer function is

$$\frac{X(z)}{F(z)} = \frac{T_D/T_1}{Z-(1-T_D/T_1)}$$

Similar expressions for other integration algorithms can be derived from the formulas given in the Appendix, and numerical stability is given by the roots of the characteristic equations of the corresponding transfer functions. The relationships between the integration step size T_D and the coefficient $1/T_1$ of the differential equation is given graphically in fig. 1. The limits of the different integration routines are given and these are important

information for the proper selection of one specific integration routine to be used. The actual regions of instability were verified by the solution of the first order differential equation on an IBM 360/50 digital computer.

Although the calculations of a problem may be numerically stable, the accuracy must also be considered, and this is dependent both on the integration step size and the forcing function. For instance quite different results were obtained by using a step function and a ramp function for the first order differential equation with the same integration time increment.

For higher order differential equations parametric expressions between the integration step size and the coefficients of the differential equation can be obtained. For instance, the second order differential equation

$$\ddot{x} + c_1 \dot{x} + c_o x = f$$

solved by means of the Rectangular Forward Integration algorithm

$$x_n = x_{n-1} + T_D \frac{dx_{n-1}}{dt}$$

yields the z-transform transfer function

$$\frac{X(z)}{F(z)} = \frac{T_D^2}{Z^2 + (C_1 T_D - 2)\, Z - C_1 T_D + C_o T_D^2 + 1}$$

Upon bilinear transformation and utilization of the Hurwitz's stability criterion the conditions for numerical stability are

$$T_D \leqslant C_1/C_o \text{ and } C_1 T_D \leqslant \frac{1}{2} C_o T_D^2 + 2$$

4. CONCLUSION

The usefulness of the closed-formula expression of numerical stability for each of the integration methods is obvious. The numerical stability for the individual integrators and the entire system can be determined from these formulas, and the best integration algorithm with appropriate time increment can be selected. This means that more than one integration algorithm and time increment can be contained in the system solution. The analysis methods would apply to both linear and non-linear systems as long as the bounds of the time-varying coefficients are known. By proper selection of the integration methods and time increments, an optimum solution can be obtained in terms of accuracy and computation time with maintained computational stability.

5. ACKNOWLEDGMENT

Gratitude is expressed to the Statistical and Computing Center of the University of Hawaii for making computer time available.

6. REFERENCES

1. Bekey, G. A. and Karplus, W. J.: "Hybrid Computation," Chap. 4. John Wiley & Sons, Inc., New York, 1968.

2. Gilbert E. G.: "Dynamic-Error Analysis of Digital and Combined Analog-Digital Computer Systems," Simulation, Vol. 6, No. 4, April, 1966, pp. 241-257.

3. Hogan, H. A. and Higgins, T. J.: "Stability Boundaries for Fifth-Order Servomechanism," Proceeding of the National Electronics Conference, Vol. XI, 1955.

4. IBM Manual E20-0029-1, "Numerical Techniques for Real-Time Digital Flight Simulation," p. 10.

5. Lapidus, L.: "Numerical Analysis for Chemical Engineers," McGraw-Hill Book Co., New York, 1962, p. 90.

6. Hildebrand, F. B.: "Introduction to Numerical Analysis," McGraw-Hill Book Co., New York, 1956, pp. 233-237.

7. Burgin, G. H.: MIDAS III ... a Compiler Version of MIDAS, Simulation, Vol. 6, No. 3, March, 1966, p. 164.

8. Fowler, M. E.: "A New Numerical Method for Simulation," Simulation, Vol. 4, No. 5, May, 1965.

9. Romanelli, M. J.: "Runge-Kutta Methods for the Solution of Ordinary Differential Equations," Chap. 9 of Mathematical Methods for Digital Computers, Edited by Ralston and Wilf, John Wiley & Sons, New York, 1966.

10. Tou, J. T.: "Digital and Sampled-Data Control Systems," McGraw-Hill Book Co., 1959, p. 244.

7. APPENDIX

Numerical Integration Formulas and their application to the differential equation:

$$\frac{dx(t)}{dt} = y(t) = \frac{1}{T_1}\left[f(t) - X(t)\right]$$

Notation: $X(t)$ = dependent variable (to be solved)

$f(t)$ = forcing function

t = independent variable, time

T_1 = system time constant

T_D = integration step size

n = index at time of integration

1 Forward Rectangular (Euler) Integration

Integration Formula: $X_n = X_{n-1} + T_D \frac{dX_{n-1}}{dt}$

Difference Equation: $X_n = (1 - \frac{T_D}{T_1}) X_{n-1} + \frac{T_D}{T_1} f_{n-1}$

2 Backward Rectangular (Euler) Integration:

Integration Formula: $X_n = X_{n-1} + T_D \frac{dx_n}{dt}$

Difference Equation: $X_n = \frac{1}{1 + T_D/T_1} X_{n-1} + \frac{T_D/T_1}{1+T_D/T_1} f_n$

3 Trapezoidal Integration

Integration Formula: $X_n = X_{n-1} + \frac{T_D}{2}\left[\frac{dx_n}{dt} + \frac{dx_{n-1}}{dt}\right]$

Difference Equation: $X_n = \frac{2-T_D/T_1}{2+T_D/T_1} X_{n-1} + \frac{2T_D/T_1}{2+T_D/T_1} f_{n-1}$

4 Huen Integration:

Integration Formula: $X_n = X_{n-1} + \frac{1}{4}(K_o + 3K_2)$

where $K_o = T_D Y(t_{n-1}, X_{n-1})$

$K_o = T_D Y(t_{n-2/3}, X_{n-1} + \frac{1}{3} K_o)$

$K_2 = T_D Y(t_{n-1/3}, X_{n-1} + \frac{2}{3} K_1)$

Difference Equation: $X_n = \left[1 - \frac{T_D}{T_1} + \frac{(T_D/T_1)^2}{2_2} - \frac{(T_D/T_1)^3}{6}\right] x_{n-1} +$

$+\left[\frac{T_D/T_1}{4} + \frac{(T_D/T_1)^3}{6}\right] f_{n-1} + \frac{(T_D/T_1)^2}{2} f_{n-2/3} + \frac{3T_D/T_1}{4} f_{n-1/3}$

5 Fowler Integration

Difference Equation: $X_n = e^{-T_D/T_1} X_{n-1} + (1-e^{-T_D/T_1}) f_{n-1}$

6 Third Order Runge-Kutta Integration

Integration Formula: $X_n = X_{n-1} + \frac{1}{6}(K_o + 4K_1 + K_2)$

6 Third Order Runge-Kutta Integration (Continued)

where, $K_o = T_D Y(t_{n-1}, X_{n-1})$

$$K_1 = T_D Y(t_{n-0.5}, X_{n-1} + \frac{1}{2} K_o)$$

$$K_2 = T_D Y(t_n, X_{n-1} + 2K_1 - K_o)$$

Difference Equation: $$X_n = \left[1 - \frac{T_D}{T_1} + \frac{2(T_D/T_1)^2}{3} - \frac{(T_D/T_1)^3}{6}\right] X_{n-1} +$$

$$+ \left[\frac{T_D/T_1}{6} - \frac{(T_D/T_1)^2}{3} + \frac{(T_D/T_1)^3}{6}\right] f_{n-1} +$$

$$\left[\frac{2T_D/T_1}{3} - \frac{(T_D/T_1)^2}{3}\right] f_{n-0.5} + \frac{T_D/T_1}{6} f_n$$

7 Fourth Order Runge-Kutta Integration

Integration Formula: $$X_n = X_{n-1} + \frac{1}{6}(K_o + 4K_1 + K_2)$$

where $K_o = T_D Y(t_{n-1}, X_{n-1})$

$$K_1 = T_D Y(t_{n-0.5}, X_{n-1} + \frac{1}{2} K_o)$$

$$K_2 = T_D Y(t_{n-0.5}, X_{n-1} + \frac{1}{2} K_1)$$

$$K_3 = T_D Y(t_n, X_{n-1} + K_2)$$

Difference Equation: $$X_n = \left[1 - \frac{T_D}{T_1} + \frac{(T_D/T_1)^2}{2} - \frac{(T_D/T_1)^3}{6} + \frac{(T_D/T_1)^4}{24}\right]$$

$$X_{n-1} + \left[\frac{T_D/T_1}{6} - \frac{(T_D/T_1)^2}{6} + \frac{(T_D/T_1)^3}{12} - \frac{(T_D/T_1)^4}{24}\right] f_{n-1} +$$

$$\left[\frac{2T_D/T_1}{3} - \frac{(T_D/T_1)^2}{3} + \frac{(T_D/T_1)^3}{12}\right] f_{n-0.5} + (T_D/6T_1) f_n$$

8 Runge-Kutta-Gill Integration

Integration Formula: $$X_n = X_{n-1} + \frac{1}{6} K_o + (2-\sqrt{2})K_1 + (2+\sqrt{2}) K_2 + K_3$$

where $K_o = T_D Y(t_{n-1}, X_{n-1})$

$$K_1 = T_D Y(t_{n-0.5}, X_{n-1} + \frac{1}{2} K_o)$$

8 Runge-Kutta Gill Integration (Continued)

$$K_2 = T_D Y\ t_{n-0.5}, X_{n-1} + (-\tfrac{1}{2} + \tfrac{1}{\sqrt{2}})\ K_o + (1 - \tfrac{1}{\sqrt{2}})K_1$$

$$K_3 = T_D Y\ t_n, X_{n-1} + (-\tfrac{1}{\sqrt{2}})K_1 + (1 + \tfrac{1}{\sqrt{2}})K_2$$

Difference Equation:

$$X_n = \left[1 - \frac{T_D}{T_1} + \frac{(T_D/T_1)^2}{2} - \frac{(T_D/T_1)^3}{6} + \frac{(T_D/T_1)^4}{24}\right]$$

$$X_{n-1} + \left[\frac{T_D/T_1}{6} - \frac{(T_D/T_1)^2}{6} + \frac{(T_D/T_1)^3}{12} - \frac{(T_D/T_1)^4}{24}\right] f_{n-1} +$$

$$+\left[\frac{T_D/T_1}{3} - \frac{(T_D/T_1)^2}{2} + \frac{(T_D/T_1)^3}{12}\right] f_{n-0.5} + \frac{T_D/T_1}{6} f_n$$

9 Runge-Kutta-Merson Integration

Integration Formula: $X_n = X_{n-1} + \frac{1}{2}\ K_o + 4K_3 + K_4$

$$K_o = \frac{T_D}{3}\ Y(t_{n-1}, X_{n-1})$$

where

$$K_1 = \frac{T_D}{3}\ Y(t_{n-2/3}, X_{n-1} + K_o)$$

$$K_2 = \frac{T_D}{3}\ Y(t_{n-2/3}, X_{n-1} + \tfrac{1}{2}K_o + \tfrac{1}{2}K_1)$$

$$K_3 = \frac{T_D}{3} Y(t_{n-0.5}, X_{n-1} + \tfrac{3}{8}K_o + \tfrac{9}{8}K_2)$$

$$K_4 = \frac{T_D}{3}\ Y(t_n, X_{n-1} + \tfrac{3}{2}K_o - \tfrac{9}{2}K_2 + 6K_3)$$

Difference Equation:

$$X_n = \left[1 - \frac{T_D}{T_1} + \frac{(T_D/T_1)^2}{2} - \frac{(T_D/T_1)^3}{6} + \frac{(T_D/T_1)^4}{24}\right.$$

$$\left. - \frac{(T_D/T_1)^5}{144}\right] X_{n-1} + \left[\frac{(T_D/T_1)^3}{8} - \frac{(T_D/T_1)^4}{48}\right] f_{n-2/3} +$$

$$+\left[\frac{2(T_D/T_1)}{3} - \frac{(T_D/T_1)^3}{3}\right] f_{n-0.5} + \frac{T_D/T_1}{6} f_n$$

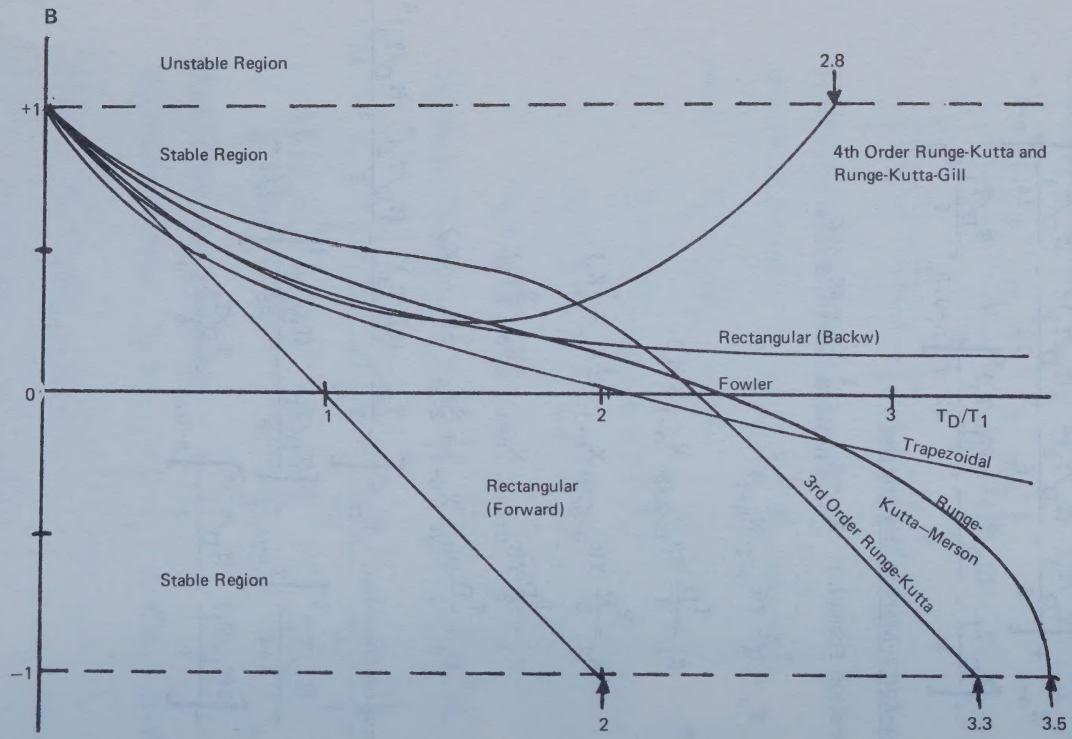


FIGURE 1

Numerical stability of a single integrator as a function of the relative time increment T_D/T_1.

Condition for stability is:

$-1 \leqslant B \leqslant 1$

For the transfer function:

$$\frac{X(z1)}{F(z)} = \frac{A}{z-B}$$

IDENTIFICATION, CLASSIFICATION, AND REDUCTION OF COMPUTATIONAL ERRORS IN THE NASAP FLOWGRAPH METHOD+

DENNIS SESAR
Hughes Aircraft Company, Culver City
L.P. McNAMEE
Computer Science Department, School of Engineering and Applied Science, University of California, Los Angeles

ABSTRACT

Computational errors in the flowgraph method of NASAP are identified and classified to their significance in the subsequent frequency response and transient response calculations. Two different implementations of the flowgraph method are described: the NASAP/ERC and the NASAP/UCLA versions. Recommendations for minimizing some of the computational errors are also presented.

1. PRELIMINARY CONSIDERATIONS

NASAP is a computer-aided circuit analysis program whose underlying philosphy differs from other such programs. It employs the flowgraph approach and solves for a desired system transfer function,[1] whereas such programs as SCEPTRE, ECAP and CIRCUS utilize a matrix state space approach to iteratively solve circuit equations for specific currents and voltages.[2]

The NASAP flowgraph method of circuit analysis requires an appropriate linear model. The model is presented to the NASAP program in terms of a linear graph whose elements are assigned to either a tree or co-tree by a hierarchial algorithm.

By performing a series of distinct operations on the linear graph, NASAP derives a particular flowgraph and then the desired system transfer function. Kirchhoff's laws are applied to generate a consistent set of voltage and current equations which are described by two disjoint flowgraphs. The two flowgraphs are then interconnected by invoking Ohm's law, establishing a complete flowgraph description of the equations. This open loop flowgraph is then closed by a branch relating the input to the output. Finally, though application of the Shannon-Happ formula the system transfer function is determined as a function of circuit parameters.[3]

The Shannon-Happ formula is defined in flowgraph notation by

$$H = 1 + \sum_{r=1}^{N} \sum_{i=1}^{R} L_i(r)\,(-1)^r = 0 \qquad (1)$$

when R= the number of r-order loops in the system, N= the order of the highest order loop, and L_i is ith loop of order i whose value is the product of all constituent elements contained in the loop.[3]

Once the system transfer function is defined, the NASAP program derives the associated AC analysis (Bode Plot) and transient response.

This paper is concerned with the computational errors associated with two versions of NASAP: NASAP/ERC developed under the direction of the NASA Electronics Research Center and NASAP/UCLA, an alternate approach developed at the University of California, Los Angeles. The implementations differ in two respects. First, the techniques to establish flowgraphs and transfer functions are fundamentally different. Second, in calculating transient responses, NASAP/ERC employs the inverse Laplace Transform method of residues utilizing the Muller rootfinding method; NASAP/UCLA uses the Fast Fourier Transform Algorithm.

For each version of NASAP, the impact and effect of errors at individual stages of the computation are presented and in some cases recommendations regarding their minimization are also included. The first part of the

+Work sponsored by NASA Research Grant NGR 05-007-200.

paper describes the errors associated with application of the Shannon-Happ formula, including Input/Output errors, significance loss, and round-off errors. These errors are manifested as deviations in computed transfer function coefficients. After the error analysis of the flowgraph approach in general is treated, the errors associated with transfer function analysis are discussed. Here, the approaches of both NASAP implementation are considered independently.

2. FLOWGRAPH ALGORITHM ERRORS

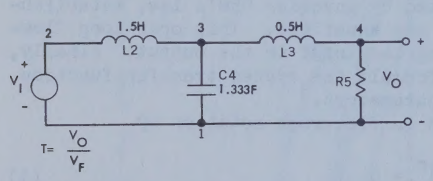
If the flowgraph algorithm and application of the Shannon-Happ formula were implemented utilizing a symbolic, or tagging approach no errors would be introduced into the derived transfer function. The NASAP programs discussed here perform all operations on numerical data. Thus, the constraint of digital computing introduces errors. Definition of these errors and how they relate to NASAP follows.

2.1 Input/Output Errors[4]

Circuit parameters are input to NASAP in decimal floating point notation. Binary arithmetic units require that the input data be converted to floating point notation compatible with the machine. Errors are introduced in program working data because there is no one-to-one correspondence between fractions of different radicies when operations are restricted to a finite wordlength. The magnitude of the error is dependent on machine hardware and the algorithm used to perform the conversion.

After required calculations are completed, the internal binary results are converted back to decimal notation. The conversion effects only the last significant digit of each coefficient and is quite minor. Also, conversion is not performed when further operations are carried out in the transfer function. For these reasons, no attempt is made here to analyze these errors.

An example of actual input/output errors (IBM 360/75) is shown in Figure 1



Data Operated Upon by NASAP Program

$L_2 = 1.49999905 \times 10^{0}$
$L_3 = 4.99999821 \times 10^{-1}$
$C_4 = 1.33299828 \times 10^{0}$
$R_5 = 1.00000000 \times 10^{0}$

Figure 1

2.2 Round-off Error

Round-off errors result directly from the finite wordlength of the machine. Errors can occur after both additive and multiplicative operations. The severe additive type errors are called significance loss, and will be described in Section 2.3.

Multiplying two numbers, each with n significant digits yields a product significant to 2n places. To store the product it is rounded to n significant digits. Wilkinson[5] has shown that in a binary rounding system the floating point computation, $Fl\,(X_1, X_2 \ldots X_n)$, produces the error bound E,

$$|E| \leq (n-1)2^{-t} + \frac{(n-1)(n-2)2^{-2t}}{2!} + \ldots \qquad (2)$$

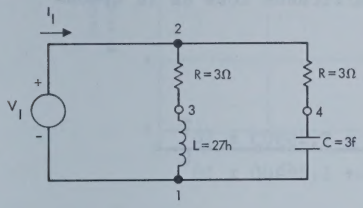
where;

$$Fl(X_1, X_2, \ldots X_n) = X_1 \cdot X_2 \ldots X_n(1+E)$$

When computing high order loops, the flowgraph technique requires a number of consecutive multiplications. The round-off errors introduced can contribute significantly to the noise contained in the coefficients of a computed transfer functions. Relation (2) shows that increasing the

wordlength, t, results in a decreasing error ratio, E. It is proposed that double precision arithmetic be utilized in carrying out all required flowgraph multiplications. The round-off noise is located in the lower order digits. In most cases, results of multiplicative operations can be guaranteed to the significance of input data.

An example of round-off error is shown in Figure 2. The errors are totally attributed to round-off error because: (1) Component values were selected such that the binary equivalent of the input data was exact. (2) Number of calculations and element values indicate no significant loss could have taken place.



Output requested: $I_I/V_I = T$

$T_{True} = I_I/V_I = 1/3$

$$T_{NASAP} = \frac{1}{3 + \left[\frac{2 \times 10^{-5}}{3S^2 + 0.667S} + 3.709 \times 10^{2}\right]}$$

Figure 2

2.3 Loss of Significance Errors

Loss of significance errors occur during floating point addition operations involving two numbers of opposite sign. The maximum error, E, resulting from the floating point addition X + Y is bounded by[5]

$$|E| \leq (1 \tfrac{1}{2})\, 2^{-t}$$

where:

$$Fl(X + Y) = (X + Y)\,(1 + E)$$

The error impact is largely dependent on the magnitude and polarity of the parameters involved, and the order in which computations are carried out. Two situations demonstrate the extent of significance loss. First consider a simple floating point addition of two numbers of opposite sign but comparable magnitudes. The resulting sum is located in the low order positions, the same locations affected by conversion and round-off errors. When the sum is left-shift normalized, the true precision of the result is disguised. The second case involves a set of repeated additions where the order of calculation can introduce loss of significance.

Significance loss errors can occur in the calculations associated with the flowgraph technique. Circuits containing elements with values covering a wide range are especially prone. The table below presents three ways in which significance loss can be detected, controlled or eliminated.

Method	Purpose
Compare results of normalized and non-normalized arithmetic	Detection
Time scaling	Minimization or Elimination
Optimum tree	Minimization or Elimination

The first approach employs a two pass system with the first pass utilizing regular floating point arithmetic.[6] During the second pass, the arithmetic is changed such that sums are not left-shift normalized. When the number is subsequently required in another calculation, the appropriate number of zeros (necessary to align the two operands) are shifted into the least significant bit positions. This scheme defines a lower bound for the calculation. The results of the two passes are examined and significance loss has occurred if these numbers are not equivalent. The significance of the calculation is considered to be equal to the number of digits which are

identical. The calculation of Figure 1 is not prone to significance loss;

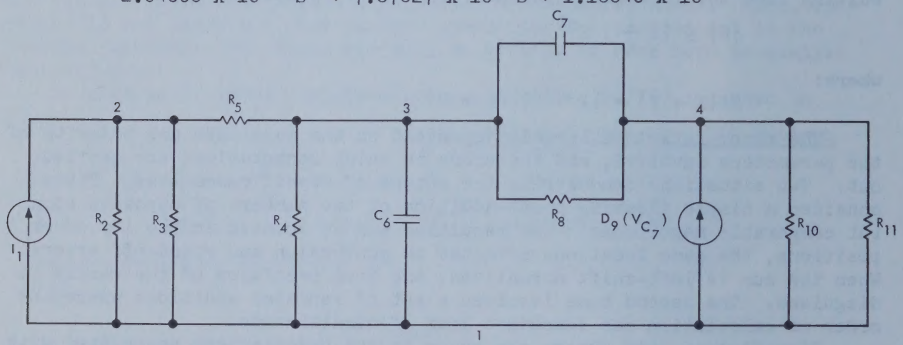
$$T = \left(\frac{V_1}{V_{R_5}}\right)_{\text{normalized}} = \frac{1.00025}{1.00000S^3 + 2.00000S^2 + 2.00050S + 1.00025}$$

$$T = \left(\frac{V_1}{V_{R_5}}\right)_{\text{unnormalized}} = \frac{1.00025}{1.00000S^3 + 2.00000S^2 + 2.00050S + 1.00025}$$

Figure 3 is a problem vulnerable to significance loss as is demonstrated below:

$$T = \frac{V_I}{I_I}_{\text{normalized}} = \frac{2.04000 \times 10^{-11}S^2 + 1.10709 \times 10^{-1}S + 7.19903 \times 10^{7}}{2.04000 \times 10^{-13}S^2 + 7.67740 \times 10^{-4}S + 1.16400 \times 10^{4}}$$

$$T = \frac{V_I}{I_I}_{\text{unnormalized}} = \frac{2.04000 \times 10^{-11}S^2 + 1.10709 \times 10^{-1}S + 7.19903 \times 10^{7}}{2.04000 \times 10^{-13}S^2 + 7.67827 \times 10^{-4}S + 1.16640 \times 10^{4}}$$



$I_1 = 1$ a
$R_2 = 1\Omega$
$R_3 = -1\Omega$
$R_4 = 100\Omega$
$R_5 = 1 \times 10^6\Omega$
$C_6 = 6$ pf
$C_7 = 200$ pf
$R_8 = 150\Omega$
$D_9 = 0.4$
$R_{10} = 1.7 \times 10^4\Omega$
$R_{11} = 100\Omega$

Compute $T = \frac{V_{R2}}{I_1}$

Figure 3

Figure 4 shows how the span of element magnitudes in Figure 3 changes as the time scale is decreased. The particular transfer function was computed and checked for significance loss at the various time scales. The results of the technique although inconclusive, tend to show that time scaling in one technique to combat significance loss.[4]

Another way of reducing significance loss in the computation of Figure 3 is to select an optimal tree. An optimal tree is defined as the tree

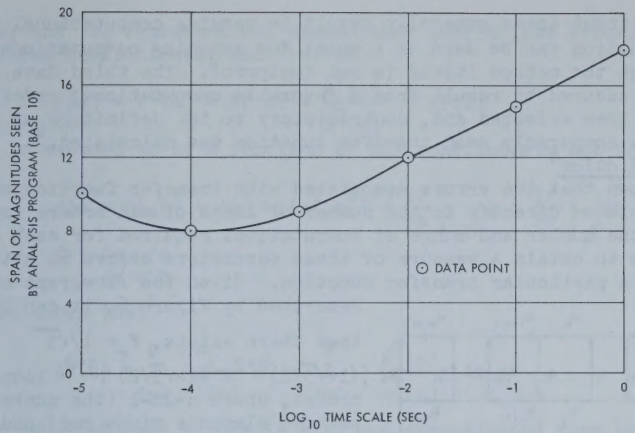


Figure 4

which, if selected, displays to the program the minimum variance between the magnitudes of the input data. The results of optimal tree selection are plotted below.

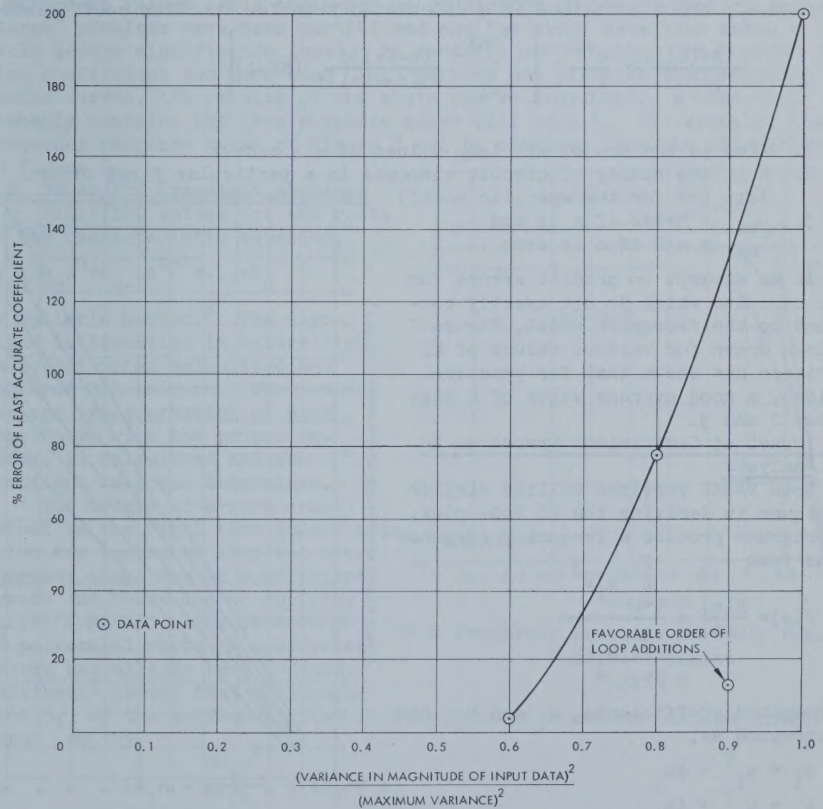


Figure 5

Different input trees generally result in varying computational orders. The above definition can be used as a means for reducing computational errors, although the method itself is not foolproof. The third data point in Figure 5 is assumed to result from a favorable computational order for the particular tree selected and, contradictory to the definition of an optimum tree, a comparably good transfer function was calculated.[4]

2.4 Error Prediction[4]

It was shown that the errors associated with transfer function coefficients are related directly to the number of loops of all orders in the flowgraph and the number and range of computations required for each loop. It is desirable to obtain a measure of these parameters before an attempt is made to derive a particular transfer function. Given the flowgraph model described by Figure 6, it can be shown that there exists, $F_n = 1/\sqrt{5}\,[(1+\sqrt{5}/2)^{n+2} - (1-\sqrt{5}/2)^{n+2}]$ loops of all orders, where n=2m-1 (the number of passive elements minus one) and F_n= Fibonacci number of order n.

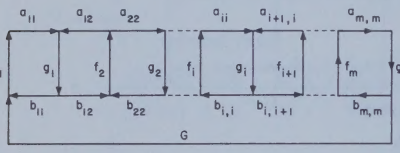


Figure 6

If flowgraph loops of all orders are assumed to have numerical values of the same magnitude, then the equation below is an estimate of the maximum round-off error contained in a calculated transfer function coefficient.

$$\frac{|\Delta\Sigma(n)|}{|\Sigma(n)|} \approx \frac{2^{-t}}{F_n} \left[\sum_{N=1}^{N_{max}} \frac{(n-N+1)!}{N!(n-2N+1)!} (NK-1) \right]$$

where

1. $\Sigma(n)$ is the sum of all loop values of all orders
2. K is the number of circuit elements in a particular first order loop (=2 for the specific model)
3. N_{max} = n+1/2 if n is odd
   N_{max} = n/2 if n is even .

In an attempt to predict errors for those circuits which do not exactly correspond to the flowgraph model, Figure 7 has been drawn for various values of K. Experience has shown that for practical problems, a good average value of K lies between 2 and 3.

3.1 Impact of Coefficient Errors on AC Analysis

Both NASAP versions utilize similar techniques in deriving the AC Bode plot. The programs produce a frequency response of the form

$$F(s) = \frac{N(s)}{D(s)} = \frac{\sum_{i=0}^{p} a_i s^i}{\sum_{j=0}^{q} b_j s^j}$$

The computed coefficients, a_i and b_j, can be expressed as,

$$a_i = a_{i_T}^{T} + \Delta a_i$$
$$b_j = b_j^{T} + \Delta b_j$$

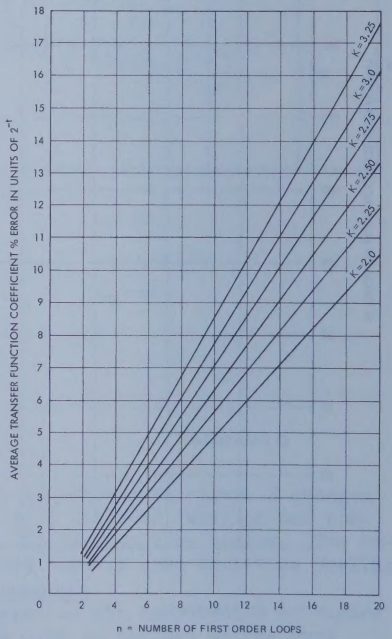

AVERAGE TRANSFER FUNCTION COEFFICIENT % ERROR IN UNITS OF 2^{-t}
n = NUMBER OF FIRST ORDER LOOPS
K=3.25
K=3.0
K=2.75
K=2.50
K=2.25
K=2.0
0 2 4 6 8 10 12 14 16 18 20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Figure 7

where a_i and b_j represent the accumulated coefficient errors a_i^T and b_i^T are the true coefficients, and F(s) as

$$F(s) = F^T(s) + \Delta F(s) = \frac{N^T(s)}{D^T(s)} + \Delta F(s) \quad .$$

If F(s) is assumed to be a function of its coefficients a,b for any s, then since[7]

$$f(a^T + \Delta a, b^T + \Delta b) \approx f(a^T, b^T) + \left.\frac{\partial f}{\partial a}\right|_{f^T} \Delta a + \left.\frac{\partial f}{\partial b}\right|_{f^T} \Delta b$$

$$\text{for } \Delta a \ll a$$
$$\Delta b \ll b \quad ,$$

it can be shown that[4]

$$\Delta F(s) = \sum_{i=0}^{p} \frac{\Delta a_i}{D^T(s)} s^i - \sum_{j=0}^{p} \frac{F^T(s)}{D^T(s)} \Delta b_j s^j \quad .$$

The computational errors in F(s) are approximated from an initial estimate of coefficient errors by substituting D(s) and F(s) for $D^T(s)$ and $F^T(s)$. These substitutions become quite laborious for functions which are a ratio of two high order polynomials. An alternate approach is taken.

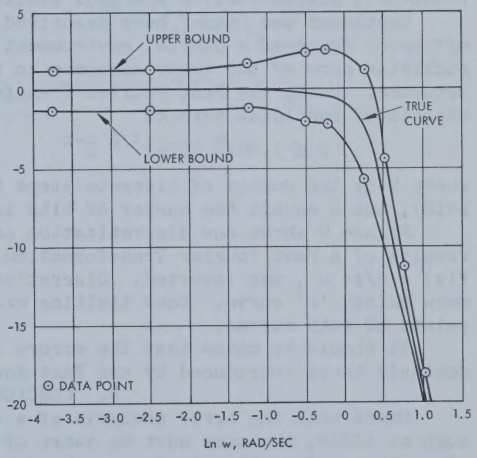
Coefficient errors are not likely to exceed 15 percent of the true coefficient values (this statement is predicated on the assumption that any "large" problems have been partitioned and the steps have been taken to avoid severe significance loss). By randomly perturbating the transfer function coefficient not more than $\pm$ 15 percent and plotting the frequency response curves, the results of the above can be simulated. A band which most probably contains the true response curve will result. For example, the frequency response curve of Figure 8 can be obtained from the circuit shown in Figure 1.

3.2 NASAP/ERC Transient Response

NASAP/ERC solves for the roots of the characteristic equation,

$$D_{(s)} = b_q s^q + b_{q-1} s^{q-1} + \ldots + b_o$$

by Muller's Method.[8] The algorithm is iterative in nature with each root being calculated and factored out in turn. The inverse Laplace transformation of each root taken with the proper residue, is calculated and the transient response determined.



True Frequency Response Curve in Band of Possible Results
Figure 8

The origin of errors contained in the final time domain solution are separated into two categories; those due to coefficient errors and those due to applying Muller's Method. The dependence of polynomial roots on coefficient errors has already been defined. Wilkinson[5] showed that a singular root S_i, of the polynomial F(s) is functionally dependent on some coefficient error Δb_k by

$$\Delta s_i \approx - \frac{\Delta b_k s^k}{F'(s_i)} \quad ; \quad \Delta s_i \ll s_i$$

When $\Delta s_i >> \Delta b_k$ the root s_i is ill-conditioned. This is the case which cause gross errors in derivation of a time domain solution from a numerically calculated transfer function.

Muller has also shown that the numerical rootfinder utilized by NASAP/ERC will converge to a true root of the particular transfer function. Using a stopping criterion of the form

$$\frac{|s_1 - s_2|}{|s_2|} < \varepsilon$$

where s_1 is the latest estimate of the desired root, and s_2 is the root calculated just prior to the estimate. Muller's Method derived roots accurate to 10^{-4} with an 80 degree polynomial. This accuracy shows that errors introduced by Muller's Method can be ignored when compared to errors due to coefficient deviations.

3.3 NASAP/UCLA Transient Response

Like Muller's Method, errors contained in a final time domain solution after application of the Fast Fourier Transform can be attributed to original coefficient errors and algorithm errors.

Given the results of Section 3.1,

$$\begin{aligned} f(t) &= \mathcal{F}[F(s)] = \mathcal{F}[F^T(s) + \Delta F(s)] \\ &= \mathcal{F}[F^T(s)] + \mathcal{F}[\Delta F(s)] \\ &= f^T(t) + \Delta f(t) \end{aligned}$$

Since the errors due to coefficient deviations are simply an additive function to the time response, an initial approximation of these errors can be transformed directly into an estimate of time response errors.

The digital implementation of the Fast Fourier Transform introduces round-off, discretization and band limiting errors.

Gentleman and Sande[9] have described the Fast Fourier Transform round-off error found in a digital environment in terms of the ratio of the Euclidian norm of the error sequence to the Euclidian norm of the input data sequence. Using the Fast Fourier Transform yield an upper bound for the associated Euclidian norm of

$$R \leq 1.06\sqrt{N}\, K(2n)^{3/2}\, 2^{-b}$$

where N is the number of discrete steps taken, $n^k = N$ (for NASAP/UCLA n=2, k=12), and b equals the number of bits in the calculation mantissa.

Figure 9 shows how discretization and band limiting errors affect the results of a Fast Fourier Transformation. For each curve, the function $F(s) = s/s + \omega$, was inverted. Discretization errors are demonstrated by the more noisey 'o' curve. Band limiting errors are shown by the poor initial points of both curves.

It should be noted that the errors resulting from coefficient errors dominate those introduced by the Fast Fourier algorithm.

4. CONCLUSION

Based upon the error analysis of a typical circuit analysis program such as NASAP, the user must be aware of the computational aspects of the algorithms before the results can be considered reliable

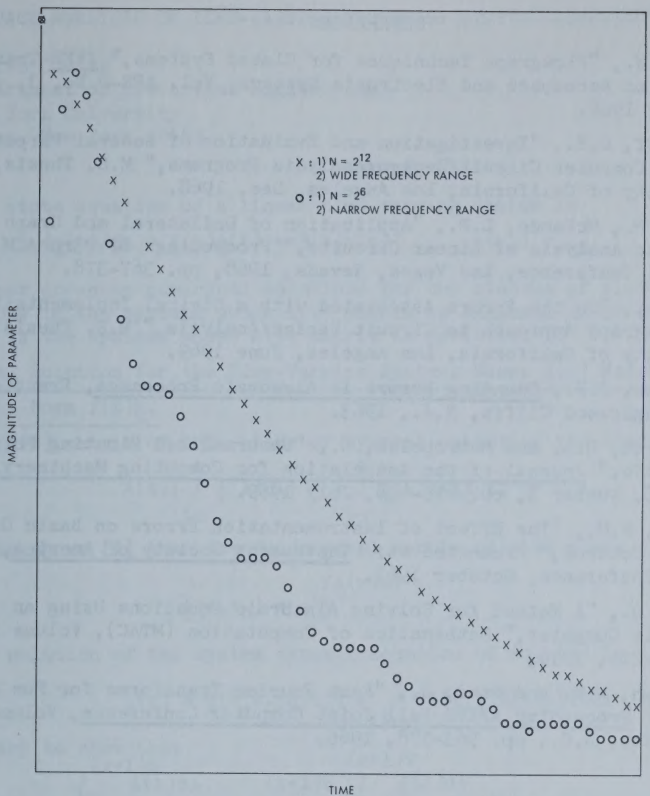

X : 1) N = 2^{12}
2) WIDE FREQUENCY RANGE
O : 1) N = 2^{6}
2) NARROW FREQUENCY RANGE
MAGNITUDE OF PARAMETER
TIME

Figure 9

REFERENCES

1. Happ, W.W., "Flowgraph Techniques for Closed Systems," IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-2 No. 3, pp. 252-264, May 1966.

2. Meyerhoff, D.E., "Investigation and Evaluation of General Purpose Digital Computer Circuit/Design/Analysis Programs," M.S. Thesis, University of California, Los Angeles, Dec. 1968.

3. Potash, H., McNamee, L.P., "Application of Unilateral and Graph Techniques to Analysis of Linear Circuits," Proceedings of 23rd ACM National Conference, Las Vegas, Nevada, 1968, pp. 367-378.

4. Sesar, D., "On the Errors Associated with a Digital Implementation of the Flowgraph Approach to Circuit Design/Analysis," M.S. Thesis, University of California, Los Angeles, June 1969.

5. Wilkinson, J.H., Rounding Errors in Algebraic Processes, Prentice-Hall Inc., Englewood Cliffs, N.J., 1963.

6. Ashenhurst, R.L. and Metropolis, N., "Unnormalized Floating Point Arithmetic," Journal of the Association for Computing Machinery, Volume 6, Number 3, pp. 415-428, July 1959.

7. Fischer, M.M., "The Effect of Instrumentation Errors on Basic Oxygen Furnace Control," Presented at: Instrument Society of America, 17th Annual Conference, October 1967.

8. Muller, D., "A Method for Solving Algebraic Equations Using an Automatic Computer," Mathematics of Computation (MTAC), Volume 10, pp. 208-215, 1956,

9. Gentleman, W.M. and Sand, G., "Fast Fourier Transforms for Fun and Profit," Proceeding AFIPS Fall Joint Computer Conference, Volume 29, Washington, D.C., pp. 563-578, 1966.

STATE SPACE ANALYSIS OF TIME-VARYING SYSTEM BY DIGITAL COMPUTER*

SAMUEL C. LEE
Department of Electrical Engineering
New York University
Bronx, New York 10453

SUMMARY

The state equation of a linear time-varying system is

$$d\underset{\sim}{x}(t)/dt = \underset{\sim}{A}(t)\underset{\sim}{x}(t) + \underset{\sim}{B}(t)x(t) \qquad \underset{\sim}{x}(t_o) = \underset{\sim}{x}_o \tag{1}$$

This paper presents numerical solutions for two classes of time-varying systems. One is the systems whose A(t) matrix is diagonal or of the form $f(t)\underset{\sim}{K}$. Another is the systems whose A(t) matrix is periodic.

Numerical Solution for the Time-Varying Systems Whose $\underset{\sim}{A}(t)$ Matrix is Diagonal or of the Form $f(t)\underset{\sim}{K}$.

The $\underset{\sim}{A}(t)$ matrix of this class of systems satisfies the condition

$$\underset{\sim}{A}(t)\left[\int_{t_o}^{t} \underset{\sim}{A}(\lambda)d\lambda\right] = \left[\int_{t_o}^{t} \underset{\sim}{A}(\lambda)d\lambda\right]\underset{\sim}{A}(t) \tag{2}$$

It can be shown [1] that the fundamental matrix of this class of systems is

$$\underset{\sim}{\Phi}(t,t_o) = e^{\int_{t_o}^{t}\underset{\sim}{A}(\tau)d\tau} \tag{3}$$

and the solution of the system dynamic equation of Eq. (1) is

$$\underset{\sim}{x}(t) = \underset{\sim}{\Phi}(t,t_o)\underset{\sim}{x}(t_o) + \int_{t_o}^{t} \underset{\sim}{\Phi}(t,\tau)B(\tau)x(\tau)d\tau \tag{4}$$

It is easy to show that

$$\underset{\sim}{x}((k+1)h) = e^{\int_{kh}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda}\underset{\sim}{x}(kT) + \int_{kh}^{(k+1)h} e^{\int_{\tau}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda}\underset{\sim}{B}(\tau)\underset{\sim}{x}(\tau)d\tau \tag{5}$$

where $t=kh$, $k=k_o, k_o+1, k_o+2, \cdots$, $t_o=k_o h$.

Evaluate $e^{\int_{kh}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda}$

The method of evaluating this matrix is briefly summarized here. It is an extension of a method due to Liou [2]. Applying Simpson's rule

$$\int_{kh}^{(k+1)h} \underset{\sim}{A}(\lambda)d\lambda = \frac{h}{6}\{\underset{\sim}{A}(kh) + 4\underset{\sim}{A}[(k+\tfrac{1}{2})h] + \underset{\sim}{A}[(k+1)h]\} - \frac{h^5}{2880}\underset{\sim}{A}^{iv}(\xi) \tag{6}$$

where $kh < \xi < (k+1)h$. Assume that h is sufficiently small, the error term of Eq. (6) is negligible. It is clear that the matrices A(kh), $A((k+\frac{1}{2})h)$, and $A((k+1)h)$ commute. Therefore

$$e^{\int_{kh}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda} \cong e^{\frac{h}{6}\underset{\sim}{A}(kh)}\, e^{\frac{4h}{6}\underset{\sim}{A}((k+\frac{1}{2})h)}\, e^{\frac{h}{6}\underset{\sim}{A}((k+1)h)} \tag{7}$$

Define

$$\underset{\sim}{M}(1) = \sum_{k=0}^{K} \frac{[\frac{h}{6}\underset{\sim}{A}(kh)]^k}{k!} \qquad \underset{\sim}{R}(1) = \sum_{k=K+1}^{\infty} \frac{\frac{h}{6}\underset{\sim}{A}(kh)^k}{k!}$$

$$\varepsilon(1) = \frac{\|\underset{\sim}{A}(kh)\|}{6(K+2)} \qquad b(1) = \frac{[\|A(kh)\|\cdot h]^{K+1}}{6^{K+1}(K+1)!}\cdot\frac{1}{1-\varepsilon(1)}$$

*The work is supported in part by the National Science Foundation Grant GK-3791.

$M^{(i)}, R^{(i)}, \varepsilon^{(i)}$, and $b^{(i)}$, i=2,3 are defined in a similar way. The procedure is given as follows: (1) Choose an initial value K arbitrarily; (2) Evaluate $\underset{\sim}{M}^{(i)}$ and then $\underset{\sim}{M}=\pi\, \underset{\sim}{M}^{(i)}$; (3) Evaluate $\underset{\sim}{M}^{(ij)}$, and then $\|\underset{\sim}{M}^{(ij)}\|$; (4) Determine $\varepsilon^{(i)}, b^{(i)}$, and b which is defined by $b=\|M^{(23)}\|\cdot b^{(1)}+\|M^{(31)}\|\cdot b^{(2)}+M^{(12)}\cdot b^{(3)}$; (5) Test each element of $\underset{\sim}{M}$ obtained from (2) with the upper bound of $|r_{ij}|$ obtained from (4) by $b\leq 10^{-d}/m_{ij}$ where d is the number of significant digits required. If it is not satisfied, then increase the value of K and repeat (1) through (4) until it is satisfied. Then

$$e^{\int_{kh}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda} \approx \underset{\sim}{M}.$$

Now turn to Eq. (5). If Simpson's rule is applied, the second integral is approximately equal to

$$\frac{h}{6}\left[e^{\int_{kh}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda}\underset{\sim}{B}(kh)x(kh)+4e^{\int_{(k+\frac{1}{2})h}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda}\underset{\sim}{B}((k+\tfrac{1}{2})h)x((k+\tfrac{1}{2})h)\right.$$

$$\left.+\underset{\sim}{B}((k+1)h)\underset{\sim}{x}((k+1)h)\right].$$

Here it is desirable to use the trapezoidal rule to evaluate the integral $\int_{(k+\frac{1}{2})h}^{(k+1)h}\underset{\sim}{A}(\lambda)d\lambda$, since the matrices $\underset{\sim}{A}((k+\frac{1}{2})h)$ and $\underset{\sim}{A}((k+1)h)$ have been calculated previously. This completes the method.

Numerical Solution for the Time-Varying Systems Whose $\underset{\sim}{A}(t)$ Matrix is Periodic

The fundamental matrix of a linear system satisfies the matrix equation:

$$\frac{d\underset{\sim}{\Phi}(t,t_o)}{dt} = \underset{\sim}{A}(t)\underset{\sim}{\Phi}(t,t_o), \quad \underset{\sim}{\Phi}(t_o,t_o) = \underset{\sim}{I} \tag{8}$$

It has been shown [3] that the solution of Eq. (8), where $\underset{\sim}{A}(t)$ is periodic with period τ and continuous for all t, has the form

$$\Phi(t,t_o) = \underset{\sim}{Q}(t,t_o)e^{\underset{\sim}{B}t} \tag{9}$$

where $\underset{\sim}{B}$ is a constant matrix, and $\underset{\sim}{Q}(t,t_o)$ has period τ. The method of evaluating $\underset{\sim}{\Phi}(t,t_o)$ is outlined below. (1) Find the numerical solution $\Phi(t,t_o)$ for $t_o\leq t\leq t_o+\tau$ using any of the known Runge-Kutter method or predictor-corrector methods. Denote it by $\Phi_1(t,t_o)$. It can be shown that $\underset{\sim}{C}=\Phi_1(t_o+\tau,t_o)$; (2) Find the eigenvalues of $\underset{\sim}{C}$ and using any of the known methods (a data condensation method, a global iteration method, or a local iteration method); (3) Obtain the matrix

$$B = \log_e C = \begin{bmatrix} \log_e L_{k_1}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & \log_e L_{k_r}(\lambda_r) \end{bmatrix}$$

$$\log_e L_{k_i}(\lambda_i) = I_{k_i}\log\lambda_i + \sum_{n=1}^{k-1}\frac{(-1)^{n-1}}{n\lambda_i^n}(L_{k_i}(\lambda_i)-\lambda_i I)^n$$

(4) Sine $\underset{\sim}{Q}(t,t_o)$ is periodic, for a given t, we can always find a t* such that $t_o\leq t^*\leq t_o+k\tau$ and $t^*=t-k\tau$, $\underset{\sim}{Q}(t^*,t_o)=\underset{\sim}{Q}(t,t_o)$. Therefore $\underset{\sim}{\Phi}(t,t_o)$ of Eq. (9) can be evaluated by $\Phi(t,t_o)=\underset{\sim}{Q}(t^*,t_o)e^{\underset{\sim}{B}t}$. The numerical computation of $e^{\underset{\sim}{B}t}$ has been described previously.

REFERENCES

1. Schwarz, R.J. and Friedlan, B., Linear Systems, McGraw-Hill Book Co., 1965, pp. 114-116.
2. Liou, M.L., "A Numberical Solution of Linear Time-Invariant Systems," Proceeding of the Third Annual Allerton Conference on Circuit and System Theory, 1965.
3. Bellman, R.* Stability Theory of Differential Equations, McGraw-Hill Book Co., New York, 1953.

DESIGN OF ASYNCHRONOUS SEQUENTIAL MACHINES WITH FAULT-DETECTION CAPABILITIES

K. C. WANG

Washington State University
Pullman, Washington

ABSTRACT

Two methods for the design of asynchronous sequential machines with fault-detection capabilities are presented. In the first method redundant outputs are added to a flow table, and in the second method it is augmented with a counter-like submachine. For each method it is shown that efficient fault-detecting sequences for checking the resulting machine operations can be constructed.

INTRODUCTION

With the advancement of the technology of micro-electronics and the increasing use of integrated-circuit modules as system components, fault diagnosis for sequential machines with terminal experiments becomes a very important and urgent problem. Examples of efforts directed toward this problem are numerous. State and machine identification experiments for synchronous sequential machines (SMs) have been studied extensively in the literature [1-3]. The design of checking experiments for SMs was considered in [4-5]. Recently, Kohavi and Lavallee [6] reported a new method for this purpose. Their main results are to show that every strongly connected SM can be modified to possess a diagnosing sequence with repeated symbols, and that efficient checking sequences can be constructed for the resulting machine. While all the studies are made for SMs only, it is not clear whether the same concept and techniques can be applied to asynchronous sequential machines.

STATEMENT OF PROBLEM

This paper considers the problem of designing AMs with fault-detection capabilities, i.e. machines whose operations can be checked by efficient terminal experiments. The experiments are assumed to be simple and preset [3]. The machines are assumed to be of the following type: fully specified, normal, Mealy model, level input-output, single input variable change, and strongly connected. The faults may cause the machine to operate incorrectly but do not increase the number of states. The objective of this

paper is to present methods for modifying a given flow table and constructing efficient checking sequences for the resulting machine.

PRELIMINARY

Let M be an AM specified by a normal flow table. A permissible input sequence is a sequence of successively adjacent inputs. Henceforth, only permissible input sequences will be considered. An input-output sequence is an input sequence together with the corresponding outputs associated with the stable states. A diagnosing sequence (DS) of M with the initial input I_1 is an input sequence $I_1 I_2 \cdots I_k$ such that the initial state of M can be uniquely determined by observing the output sequence $z_1 z_2 \cdots z_k$. M is diagnosable if it has a DS for each possible initial input. M is definitely diagnosable (DD) if every input sequence of length k is a DS, where k is the order of diagnosability.

To determine the diagnosability of M, we can construct its successor tree or implication graph. The successor tree of M is defined as follows: The zeroth level consists of all the states that are stable in some column. The branches leading to the first level are labeled with the inputs; each input I_j leads to the stable state in the column of I_j, and the states are divided into groups according to their outputs. Similarly for the succeeding levels. A branch becomes terminal if (i) it leads to a group of states containing two or more identical states, or (ii) it lead to a set of states already appeared in a preceding level under the same input, or (iii) it leads to a set consisting entirely of single states. Similar to SM case, the branches terminated by rule (iii) are the DS of M.

The implication graph of M contains essentially the same information as that of the successor tree. To construct the graph, pairs of states under the same input and having the same outputs are entered first as nodes. There is a branch from one node to the other only if (i) under an adjacent input the pair of states goes to another pair of states having the same output, or (ii) the pair of states goes to the same state. In the latter case, a new node consisting of a single state is added, which is also terminal. Based on the implication graph, it is easy to see that M is DD if and only if the graph contains no single state node, and no loop. If M is DD, the order k can also be read from the graph.

Let #S be the maximal number of stable states in any column of M. Then, at most $K = \lceil \log_2 \#S \rceil$ additional output variables are needed to distinguish the states in any column. It follows that any M can be made to be DD of order 1. Similarly, any M can be made to be DD of order ℓ by adding redundant outputs and state-spliting. The latter is necessary when ℓ is to be > 1 and there exist two or more states going into a single state.

FIRST METHOD

Our first method of designing diagnosable AMs is presented as the following steps:

Step 1. Let M have 2^m inputs; each input is coded by an m-tuple. Check the diagnosability of M. If M is not DD, augment the flow table so that the resulting M is DD of order ℓ.

Step 2. Since M is normal, a single transition time (STT) state assignment can be made for M. This assignment technique ensures that each state has a unique coding and that no new state is created to achieve the transitions [7-8]. Consequently, we can construct the checking sequences based entirely on the flow table of M.

Step 3. From M construct the extended flow table $\bar{M}$. The states of $\bar{M}$ are the stable states of M; each state in $\bar{M}$ is a triple (s_i, I_j, z_k), and (I_j, z_k) will be used as the output associated with the state. The transitions of $\bar{M}$ follow those of M in a natural way, i.e. there is a transition from (s_i, I_j, z_k) to (s_i', I_j', z_k') iff such a transition exists in M. Now, $\bar{M}$ can be regarded as a synchronous (Moore) machine with m inputs $x_1, x_2, .., x_m$, where x_j stands for "change the jth component of the input in M".

Theorem: If M is DD of order ℓ, so is $\bar{M}$ when regarded as a SM.

Proof: From the way $\bar{M}$ is constructed, it takes ℓ inputs to distinguish all the states if it requires at least ℓ inputs for M to distinguish all the stable states.

Step 4. Based on the theorem, the design method in [6] can be applied here to construct the checking sequences for $\bar{M}$. The procedures are listed below:

a). Start $\bar{M}$ from an initial state. Apply a DS with repeated symbols, say $(x_1 x_1 .... x_1)$ to verify the initial state. Add a x_1 to check the transitions until no new

transitions can be generated, then go to b).

b). Apply a $x_j = x_1$ and repeat a). Repeat a) and b) until no new transitions can be generated, then go to c).

c). Apply a transfer sequence to bring $\bar{M}$ through the checked transitions to a state with unchecked transitions and repeat a),b), and c) until all transitions are checked.

The checking sequence for M is given by the output sequence in the checking sequence of $\bar{M}$. The validity of this checking sequence can be justified by the following argument:

M is operating correctly if and only if its extended flow table is identical or isomorphic to $\bar{M}$. Since the transitions of $\bar{M}$ are checked by the same DS sequence, the resulting checking sequence uniquely determines $\bar{M}$ (up to isomorphism). [6] Therefore, M is operating correctly if and only if it is consistent with the checking sequence thus obtained.

Example: Consider the flow table M_1 shown in Figure 1.

	00	01	11	10
1	(1) 0	2 0	(1) 0	4 1
2	3 1	(2) 0	1 0	(2) 0
3	(3) 1	4 0	(3) 1	2 0
4	1 0	(4) 0	3 1	(4) 1

Figure 1. M_1 ($\ell = 2$).

The extended flow table $\bar{M}_1$ is shown in Figure 2.

	x_1	x_2	t(I,z)
1	8	3	000
2	7	4	001
3	5	2	010
4	6	1	010
5	3	8	110
6	4	7	111
7	2	5	100
8	1	6	101

Figure 2. $\bar{M}_1$.

A checking sequence for $\bar{M}_1$ starting from the state s_1 is constructed and is shown as follows:

1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 2 1

1 8 1 8 1 3 5 3 5 3 2 7 2 7 2 4 6 4 6 4 1 8

1 1 2 1 1 2 1 1 2 1 1

1 8 6 4 6 7 2 7 5 3 5

The checking sequence for M can be obtained by replacing the states in the above sequence with the outputs in Figure 2.

SECOND METHOD

Our second method is to imbed a counter-like submachine into a given flow table, and to use the counter transitions to check the machine operations. The basic assumption here is that errors only occur in transitions involving the change of state variables.

A counter M_c is shown in Figure 3.

I_c	I_d
(1) 1	2 0
3 0	(2) 0
(3) 0	..
. .	. .
	2k 0
1 1	(2k) 0

Figure 3. M_c.

It is easy to verify that M_c is DD of order k+1. For any given M, we can imbed M_c into M such that each I_j of M is adjacent to both I_c and I_d. This is always possible by introducing two additional inputs. For instance, let M have two input variables. The inputs of the augmented flow table are shown by the map of Figure 4.

	00	01	11	10
00	I_1	I_2	I_3	I_4
01	I_c	I_d	I_c	I_d
11	-	-	-	-
10	I_d	I_c	I_d	I_c

Figure 4. Map of the inputs.

In the augmented machine, it is possible to bring the machine from any stable state into a stable state in the counter. Since the counter is DD, every transition can be checked in the counter. It is easy to design the checking sequences for the resulting

machine, and this is omitted here.

CONCLUSION

In conclusion, we have presented two methods for the design of AMs with fault-detection capabilities. The first method utlizes the results developed for SMs, and provides a set of procedures for constructing the checking sequences in a systematical manner. The second method yields more restricted results but it is applicable even if M is not strongly connected.

REFERENCES

1. E. F. Moore, "Gedanken-experiments on sequential machines", in Automata Studies, Princeton Univ. Press, Princeton, N.J.;1956.
2. S. Ginsburg, "On the length of the smallest uniform experiment which distinguishes the terminal states of a machine", JACM pp.226-280.1958.
3. A. Gill, Introduction to the theory of finite state machines, McGraw-Hill Co., 1962.
4. F. C. Hennie, "Fault-detecting experiments for sequential circuits", Proc. 5th Symp. Switching Theory and Logical Design, Princeton Univ.; pp.95-110; 1964.
5. C. R. Kime, "A failure detection method for sequential circuits" Dept. Elect. Engr. Univ. Iowa, Technical Report 66-13, 1966.
6. Z. Kohavi and P. Lavallee, " Design of sequential machines with fault-detection capabilities", IEEE Trans. Computers, EC-16, pp.473-484; 1967.
7. C. N. Liu, "A state variable assignment method for asynchronous sequential switching circuits", JACM pp.209-216, 1963.
8. J. H. Tracey, "Internal state assignment for asynchronous sequential machines", IEEE Trans. Electronic Computers, EC-15, pp.551-560, 1966.

RELIABILITY MODELING AND ANALYSIS OF A DYNAMIC TMR SYSTEM UTILIZING STANDBY SPARES*

FRANCIS PARKASH MATHUR
Jet Propulsion Laboratory
Pasadena, California

ABSTRACT

A dynamic TMR (triple modularly redundant) system utilizing standby spares, referred to as the TMR/Spares system in this paper, is described, modeled, and analyzed. The characteristic reliability equation of such a system is derived. The behavior of the system reliability as a function of the nonredundant system and normalized time is plotted and comparisons made against conventional TMR systems. The TMR/Spares model is considered to be a major addition to the redundancy techniques available to the designer of fault-tolerant computers.

I. INTRODUCTION

The use of protective redundancy to enhance reliability[1, 2] -- once every step has been taken, under the limitations of the prevailing state of technology, to select, screen, and package highly reliable components -- has, as a result of the research conducted and applications made in this field over the last decade,[3,4] found wide acceptance as a fundamental procedure and is a process which nature in her apparent working sanctions.[5]

Protective redundancy has been classified as being either massive or selective and also as being static or dynamic. The first classification springs from an organizational viewpoint,[2] the latter from a functional viewpoint.[6] This paper presents a system concept that is a hybrid, utilizing both static and dynamic redundancy. It is a combination of two well-known techniques, namely, an amalgamation of triple modular redundancy (TMR) and replacement system (RS) redundancy.

II. THE TMR/SPARES SYSTEM

The simple TMR types of systems are first reviewed and are illustrated in Fig. 1. A simplex or nonredundant system having reliability R or R(T) is shown in Fig. 1(a). (The term reliability throughout this paper is defined as the probability that a system required to perform a certain task will not malfunction prior to its mission time T.) The reliability of a simple TMR system as shown in Fig. 1(b) is given by the following well-known equation:

$$R(TMR) = R^3 + 3R^2(1 - R) \qquad (1)$$

*This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. NAS 7-100, sponsored by the National Aeronautics and Space Administration. A preliminary version of this paper was presented as a working paper ("Design and reliability analysis of 'hard cores' of fault-tolerant computers") at the ACM Workshop on Hardware-Software Interaction for System Reliability and Recovery in Fault-Tolerant Computers, Pacific Palisades, California, July 14-15, 1969.

The generalization of the TMR concepts[7] to a system utilizing $\eta = 2N + 1$ units and having a $N + 1/\eta$ restoring organ is illustrated in Fig. 1(c) and is designated as the NMR system; its well-known reliability equation is

$$R(NMR) = \sum_{i=0}^{N} \binom{\eta}{i} (1 - R)^i R^{\eta - i} \qquad (2)$$

The family of curves illustrating its behavior is shown in Fig. 2, with reliability plotted as a function of normalized time λT.

The underlying failure law will herein be assumed to be exponential.[8] Thus the simplex reliability R is given by $\exp(-\lambda T)$, where λ is the failure rate of the nonredundant system when it is active. In the ensuing development of the probabilistic model for the TMR/Spares systems, the assumption of statistical independence of failures has been made.

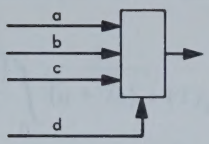
The TMR/Spares system concept (Fig. 3) consists of a TMR core, with an associated bank of S spare units such that when one of the basic TMR units fails, the spare unit replaces it and restores the TMR core to the all-perfect state. The active TMR units have a failure rate designated by λ, while the standby spare units, which are said to be in a dormant mode,[9] have a failure rate designated by $\mu (\mu \leq \lambda)$, with the corresponding reliability $R_S = \exp(-\mu T)$. The physical realization of such a system is shown in Fig. 4, where the disagreement detector (DD) compares the system output with each one of the basic triplicated units. When a disagreement occurs, a signal is transmitted to the switching unit (SU), which replaces the unit that disagreed by switching it out and switching in one of the spares. If the spare fails while in the dormant mode, the disagreement will still exist and the SU will again replace it by one of the spares. The TMR/Spares system reduces to a simple TMR system when all the spares have been exhausted, and the whole system fails upon the exhaustion of all the spares and the failure of any two of the basic triplicated units.

The TMR/Spares system concept has been considered by other researchers from the architectural standpoint.[10,11] A simple derivation[12] of the reliability equation when dormancy of the spare units is not considered (i.e., when all the S + 3 units in the system are considered to have identical failure rates) yields

$$R(TMR/S) = 1 - (1 - R)^{S+2} [1 + R(S + 2)]$$

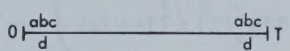
which is simply the probability that at least any two of the total S + 3 units survive the mission duration.

An expression for the reliability of TMR/1 system (i.e., a TMR/Spares system with one spare) is now derived. Let the three basic triplicated units be designated as a, b, and c, and the spare as d, as follows:



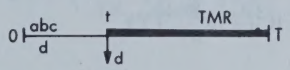
Three cases may be distinguished which yield the success of the system for any mission time T.

Case (i). All units survive time T:



This event has the probability $R^3 R_s$.

Case (ii). The spare unit is the first to fail:

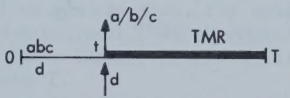


At some time t the spare unit d fails, leaving the system in basic TMR for the unelapsed time $[T - t]$. The probability of this event is

$$\int_0^T e^{-3\lambda t} \mu e^{-\mu t} R(TMR/0)[T - t]\, dt$$

where $R(TMR/0)[T - t] = R^3[T - t] + 3R^2[T - t](1 - R[T - t])$, which is the reliability of a simple TMR system for a mission time $[T - t]$.

Case (iii). An active unit fails before the spare:



At some time t either a or b or c fails and is replaced by the spare d, thus leaving the system in basic TMR for the rest of the time $[T - t]$. The probability of this event is:

$$3\int_0^T \lambda e^{-\lambda t} e^{-\mu t} e^{-2\lambda t} R(TMR/0)[T - t]\, dt$$

Summing the above three cases yields

$$R(TMR/1)[T] = R^3(T)R_s(T) + (3\lambda + \mu)\int_0^T e^{-(3\lambda+\mu)t}\, R(TMR)[T - t]\, dt \tag{3}$$

Similarly, it may be shown that

$$R(TMR/2)[T] = R^3(T)R_s^2(T) + (3\lambda + 2\mu)\int_0^T e^{-(3\lambda+2\mu)t} R(TMR/1)[T - t]\, dt \tag{4}$$

and, in general,

$$R(TMR/S+1)[T] = R^3(T)R_s^{S+1}(T) + [3\lambda + (S+1)\mu]\int_0^T e^{-[3\lambda+(S+1)\mu]t}\, R(TMR/S)[T - t]\, dt \tag{5}$$

which may be rewritten by making the appropriate substitution as

$$R(TMR/S+1)[T] = R^3(T)R_s^{S+1}(T)\left\{1 + [3\lambda + (S+1)\mu]\int_0^T e^{+[3\lambda+(S+1)\mu]\tau} R(TMR/S)[\tau]\, d\tau\right\} \tag{6}$$

The above recursive equation has the solution

$$R(TMR/S) = 3R^2\left\{\prod_{i=0}^{S-1}\left(\frac{3K+S-i}{K+S-i}\right) - \frac{2RK^2}{S!}\left[\prod_{i=0}^{S-1}(3K+S-i)\right]\sum_{i=0}^{S}\binom{S}{i}\frac{(-R_s)^{S-i}}{(K+S-i)(3K+S-i)}\right\} \tag{7}$$

where $K = \lambda/\mu$; $\mu \leq \lambda$; $1 \leq K < \infty$; and

$$\binom{S}{i} = \frac{S!}{(S-i)!\, i!}$$

For the special situation of $K = \infty$, i.e., $\mu = 0$, the solution (7) reduces to

$$R(TMR/S) = 3^{S+1} e^{-2\lambda T} - e^{-3\lambda T}\left[\sum_{i=0}^{S} \frac{(3\lambda T)^i}{i!}\left(3^{S+1-i} - 1\right)\right] \tag{8}$$

An alternative way of writing Eq. (7) is

$$R(TMR/S) = 3R^2\left[\frac{\binom{3K+S}{3K}}{\binom{K+S}{K}} - 2RK^2\binom{3K+S}{3K}\sum_{i=0}^{S}\binom{S}{i}\frac{(-R_s)^{S-i}}{(K+S-i)(3K+S-i)}\right] \tag{9}$$

The proof that the Eq. (9) is the solution to the recursive integral Eq. (6) can be verified by inserting (9) on the right-hand side of (6).

In the derivation of the above equations it was assumed that the restoring organ, the switching unit, and the disagreement detector (jointly referred to as the R-S-D unit) are perfect. In order to incorporate the unreliability of these units, they may be assigned a lumped parameter R_V, reflecting their reliability; and with the simplifying assumption that the R-S-D unit has a series reliability relative to the ideal TMR/spares configuration, the term R_V may be used as a product term to directly modify Eqs. (7), (8), and (9).

The behavior of Eq. (9) is shown graphically in Fig. 5 for the case where $K = 1$. Reliability is plotted versus the nonredundant unit, i.e., the simplex system having reliability R. A family of curves for the number of spares equal to 6, 4, 2, and 1 is shown; also, for comparative purposes, the reliability curves of the generalized TMR system (NMR) as given by Eq. (2) are also shown. The same family of curves is also shown in Fig. 6 as a function of normalized time T.

III. CONCLUSION

The TMR/Spares system has been described, and its characteristic reliability equation was derived and analyzed under the assumptions of statistically independent failures and an exponential failure law and under conditions of dormant failure rates for the spares. The behavior of the equations as a function of the nonredundant system and normalized time has been shown graphically. From these curves, the improvement in reliability resulting from the hybridization of simple TMR with RS redundancy is readily seen. It is to be noted that the well-known crossover point, which in NMR systems occurs at a reliability of 0.5, with the resulting manyfold constraints in practical implementation of such systems, is, in the TMR/Spares system, by the use of only one spare, reduced to a maximum value of 0.233 and may be further reduced by allocating more spares. Another

major advantage in terms of hardware and cost of the TMR/Spares system over the NMR system is that for an equal number of total units η the NMR system will tolerate failures of only $\eta/2$ units whereas the TMR/spares system will tolerate $\eta - 2$ failures.

IV. ACKNOWLEDGMENTS

The author wishes to thank William F. Scott, John J. Wedel, and George R. Hansen of the Guidance and Control Division of the Jet Propulsion Laboratory for their encouragement and for providing the atmosphere conducive to this research. Thanks are due Dr. Ho Sen Lin for his many helpful suggestions regarding the mathematics given here. I am deeply grateful to Prof. Algirdas A. Avizienis for introducing me to the subject of fault-tolerance and for the skillful and generous guidance given in negotiating the abundance of literature available in this field.

V. REFERENCES

1. Moore, E. F., and Shannon, C. E., "Reliable circuits using less reliable relays," J. of the Franklin Institute, 1956, Vol. 262, Pt. I, pp. 191-208, and Vol. 262, Pt. II, pp. 281-297.

2. Avizienis, A. A., "Design of fault-tolerant computers," AFIPS Conference Proc., Vol. 31, Fall Joint Computer Conference, 1967, pp. 733-743.

3. Avizienis, A. A., Mathur, F. P., Rennels, D., and Rohr, J., "Automatic maintenance of aerospace computers and spacecraft information and control systems," Proc. of the AIAA Aerospace Computer Systems Conference, Paper 69-966, Los Angeles, Sept. 8-10, 1969.

4. Anderson, J. E., and Macri, F. J., "Multiple redundancy applications in a computer," Proc. 1967 Annual Symposium on Reliability, Washington, 1967, pp. 553-562.

5. von Neumann, J., "Probabalistic logics and the synthesis of reliable organisms from unreliable components," in Automata Studies, Princeton University Press, Princeton, N. J., 1956, pp. 43-98.

6. Short, R. A., "The attainment of reliable digital systems through the use of redundancy — a survey," Computer Group News, March 1968, pp. 2-17.

7. Knox-Seith, J. K., "Improving the reliability of digital systems by rudundancy and restoring organs," Ph.D. thesis, Electrical Engineering, Stanford University, August 1964.

8. Drenick, R. F., "The failure law of complex equipment," J. Soc. Ind. Appl. Math., Vol. 8, No. 4, Dec. 1960, pp. 680-690.

9. Mathur, F. P., "Reliability study of fault-tolerant computers," in Supporting Research and Advanced Development, Space Programs Summary 37-58, Vol. III, pp. 106-113, Jet Propulsion Laboratory, Pasadena, Calif., Aug. 31, 1969.

10. Goldberg, J., Levitt, K. N., and Short, R. A., "Techniques for the realization of ultra-reliable spaceborne computers," Final Report — Phase I, Project 5580, Stanford Research Institute, Menlo Park, Calif., Oct. 1967.

11. Goldberg, J., Green, M. W., Levitt, K. N., and Stone, H. S., "Techniques for the realization of ultrareliable spaceborne computers," Interim Scientific Report 2, Project 5580, Stanford Research Institute, Menlo Park, Calif., Oct. 1967.

12. Roth, J. P., Bouricius, W. G., Carter, W. C., and Schneider, P. R., "Phase II of an architectural study for a self-repairing computer," International Business Machines Corporation, Report SAMSO TR-67-106, Nov. 1967.

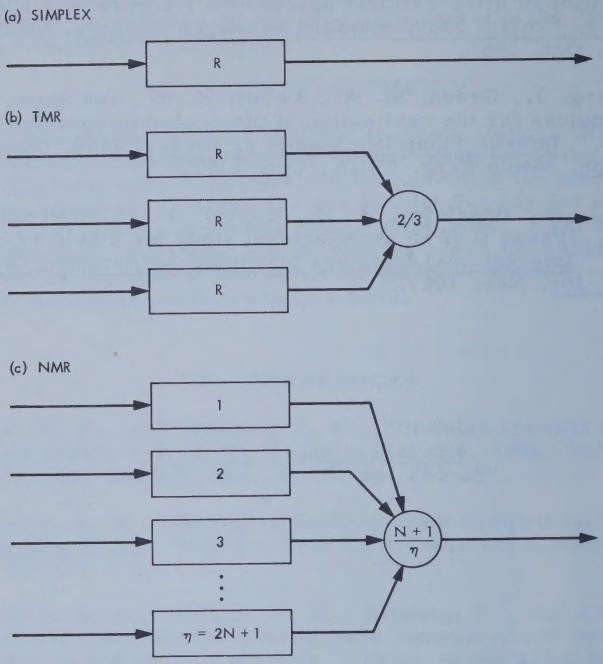


Fig. 1. Simple TMR-type systems

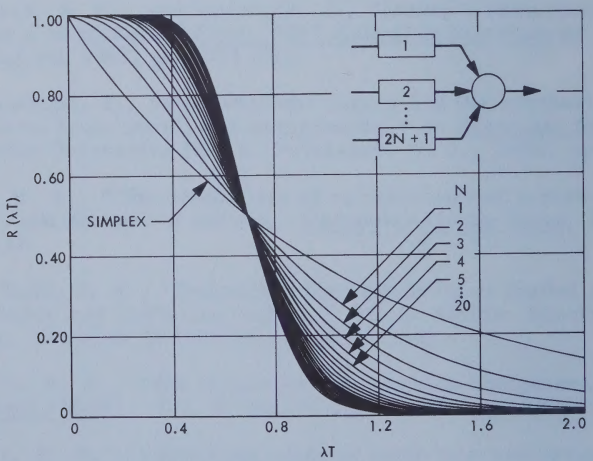


Fig. 2. Reliability of NMR-type systems vs normalized time

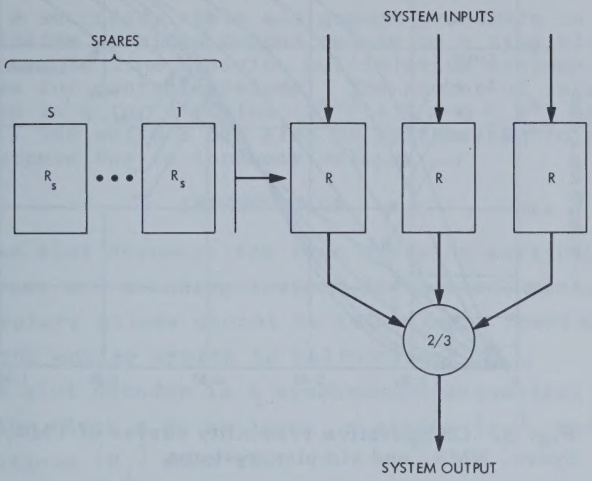


Fig. 3. TMR/Spare system concept

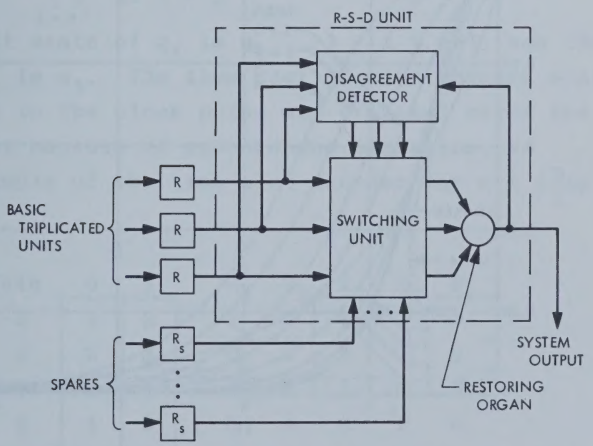


Fig. 4. TMR/Spare system block diagram

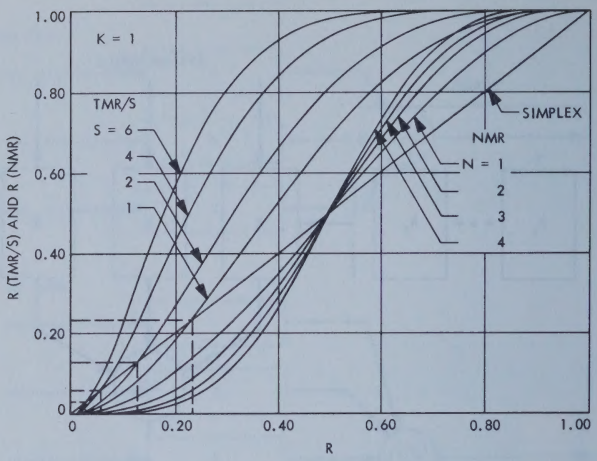


Fig. 5. Comparative reliability curves of TMR/Spare, NMR, and simplex systems

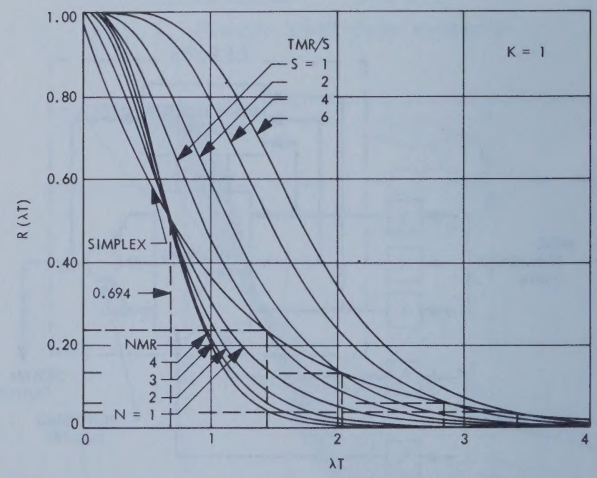


Fig. 6. Reliability comparison of a TMR/Spare system vs normalized time with and without spares

ELIMINATION OF SPURIOUS PULSES FOR TIME SLOT DECODERS

JON G. BREDESON AND PAUL T. HULINA
Pennsylvania State University
University Park, Pennsylvania

Abstract - A secondary state assignment procedure is presented which eliminates spurious output pulses on a time slot decoder which are used in time division multiplexing systems and scanning devices for control systems. The number of internal variables needed is s for n states ($2^{s-1} + 1 \leq n \leq 2^s$) except when $n = 2^s - 1$. The network can also be synthesized to eliminate spurious outputs due to inherent delays.

INTRODUCTION

In time slot decoders for time division multiplex communication systems and scanning devices for control systems spurious or momentary pulses cannot be tolerated. Spurious pulses may cause the entire system to malfunction.

A time slot decoder is a synchronous sequential machine with only the clock c as an input, n states $\{q_i\}$, and n level output functions $\{Z_j\}$ such that

$$Z_j(q_i) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad 1 \leq i, j \leq n$$

and the next state of q_i is q_{i+1}, $1 \leq i \leq n-1$, and the next state of q_n is q_1. The time slot decoder changes state only in response to the clock pulse and critical races are assumed not to exist because of synchronous operation.

An example of the time slot decoder for n = 5 is given in Fig. 1.

State	c = 0	c = 1	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	2	1	0	0	0	0
2	2	3	0	1	0	0	0
3	3	4	0	0	1	0	0
4	4	5	0	0	0	1	0
5	5	1	0	0	0	0	1

Fig. 1. Time slot decoder for n = 5

Let us use the following state assignment:

	y_1	y_2	y_3
1	0	0	0
2	0	0	1
3	0	1	1
4	0	1	0
5	1	1	0

With this assignment the machine in sequencing from state 5 to 1 can assume momentarily the internal states 100 or 010 due to the racing of y_1 and y_2. If 010 occurs, then Z_4 will momentarily become 1 during the transition. This situation cannot be tolerated because our device could be sampling two functions at once.

FUNCTION HAZARDS

Before the state assignment problem can be attempted some definitions are in order. An input state for any function will be assumed to be a function of an n-tuple $(y_1, y_2, ---, y_n)$.

Definition 1 - An input state A is said to be in the subcube of input states B and C (written $A \epsilon (B,C)$) if and only if the logic value of the y_i variable of A is the same as the value of the y_i variable of B and C whenever the logic value of B and C are the same.

As an example of Definition 1 consider the following input states:

A = (0,1,1,1) B = (0,0,1,1) C = (0,1,0,1)

By the definition $A \epsilon (B,C)$ but $C \notin (A,B)$.

Definition 2 - A Boolean function f contains a static function hazard for the input change A to C if and only if

(1) $f(A) = f(C)$
(2) There exists at least one input state $B \epsilon (A,C)$ such that $f(B) \neq f(C)$.

The previous definition is equivalent to Eichelberger's [1] definition of a function hazard.

Definition 3 - A Boolean function f contains a dynamic function hazard for the input change A to D if and only if

(1) $f(A) \neq f(D)$
(2) There exists at least one pair of input states B and C such that
    (a) $B \epsilon (A,D)$ and $C \epsilon (B,D)$ and
    (b) $f(B) = f(D)$ and $f(C) = f(A)$.

Consider the function in Fig. 2.

z \ x y	00	01	11	10
0	0^A	1^B	1	0
1	1	0^C	1^D	1

Fig. 2. Example of Function Hazards

A static function hazard exists for input change A to C and a dynamic function hazard exists for input change A to D. It should be apparent that if a function hazard exists (either static or dynamic), a sequence of input changes can exist where a function may have a multiple (more than one) output change. Then it is possible that delays in the network realizing a function with function hazards could cause a multiple change in the output of the network. It should be obvious that function hazards cannot be tolerated on the output functions of a time slot decoder for allowable internal variable changes.

THE STATE ASSIGNMENT OF TIME SLOT DECODERS

As defined in the Introduction a time slot decoder has n states $\{q_i\}$ and n outputs $\{Z_j\}$ such that $Z_j(q_i) = 1$ if and only if $i = j$. Each state q_i will be assigned an internal state s_i. Then a function Z_j^* of the internal secondary states must be formed such that $Z_j^*(s_i) = Z_j(q_i)$ when s_i is assigned to q_i. For each transition q_i to q_{i+1} we cannot tolerate function hazards on Z_j^* for input change s_i to s_{i+1}.

The following definition by Tracey [2] will be needed for what is to follow:

Definition 4 - A partition π on a set S is a collection of subsets of S such that their pairwise intersection is the null set. The disjoint subsets are called the blocks of π. If the set union of these subsets is S, the partition is completely specified, otherwise the partition is incompletely specified. Elements of S that do not appear in it are called unspecified with respect to that partition. A y-variable y_i induces the partition $\pi = \{s_1, s_2, -- s_m; s_{m+1}, ---, s_n\}$ if the value of y_i is "0" for $s_1, s_2, -- s_m$ and "1" for $s_{m+1}, --, s_n$ or vice versa. $s_i = s_j(\pi)$ means s_i and s_j are in the same block of partition π.

We will assume $Z_j^*(s)$ equals 1 only when $s = s_j$. This implies all internal states which do not correspond to states q_i will have output 0. Theorem 1 gives the necessary and sufficient conditions for assigning the internal states to avoid function hazards.

Theorem 1 - The output function Z_j^* for a time slot decoder contains no static function hazard for internal state change s_i to s_{i+1} (state transition q_i to q_{i+1}) if and only if there exists at least one internal y-variable which partitions the pair of internal states $\{s_i, s_{i+1}\}$ and the internal state s_j into separate blocks ($i \neq j \neq i+1$).

Proof - Assume Z_j^* contains no static function hazard for internal state change s_i to s_{i+1} which implies $s_j \notin (s_i, s_{i+1})$. By Definition 1, $s_j \notin (s_i, s_{i+1})$ implies there exists a y-variable which has the same value for s_i and s_{i+1} and a different value for s_j and hence partitions $\{s_i, s_{i+1}\}$ from s_j.

Assume there exists a y-variable y_k which partitions s_j from $\{s_i, s_{i+1}\}$. Thus $s_j \in (s_i, s_{i+1})$. Since s_j is the only state such that $Z_j^*(s_j) = 1$, Z_j^* can have no static function hazard for the transition s_i to s_{i+1}. Hence the theorem.

Since we assumed Z_j^* equals 1 only for internal state s_j, Z_j^* has no dynamic function hazard for any transition by Defn. 3. We must insure each output function Z_j^*(j = 1, 2, -- n) has no function hazards for any transition. Thus at least one y-variable must partition each pair of states $\{q_i, q_{i+1}\}$ from state q_j. The problem now reduces to finding a minimum set of partitions of the type $\pi = \{\overline{q_i, q_{i+1}}; \overline{q_j}\}$. The procedure for finding such partitions is exactly the same as outlined by Tracey [2], and will not be given here.

The π partitions, τ-partitions (found by Tracey's method) and state assignment are given for the time slot decoder in Fig. 1.

Z_1: $\pi_1 = \{\overline{1}; \overline{2,3}\}$ $\quad \pi_2 = \{\overline{1}; \overline{3,4}\}$ $\quad \pi_3 = \{\overline{1}; \overline{4,5}\}$

Z_2: $\pi_4 = \{\overline{2}; \overline{3,4}\}$ $\quad \pi_5 = \{\overline{2}; \overline{4,5}\}$ $\quad \pi_6 = \{\overline{2}; \overline{5,1}\}$

Z_3: $\pi_7 = \{\overline{3}; \overline{4,5}\}$ $\quad \pi_8 = \{\overline{3}; \overline{5,1}\}$ $\quad \pi_9 = \{\overline{3}; \overline{1,2}\}$

Z_4: $\pi_{10} = \{\overline{4}; \overline{5,1}\}$ $\quad \pi_{11} = \{\overline{4}; \overline{1,2}\}$ $\quad \pi_{12} = \{\overline{4}; \overline{2,3}\}$

Z_5: $\pi_{13} = \{\overline{5}, \overline{1,2}\}$ $\quad \pi_{14} = \{\overline{5}; \overline{2,3}\}$ $\quad \pi_{15} = \{\overline{5}; \overline{3,4}\}$

$\tau_1 = \{\overline{1,2}; \overline{3,4,5}\}$

$\tau_2 = \{\overline{1,5}; \overline{2,3,4}\}$

$\tau_3 = \{\overline{2,3}; \overline{1,4,5}\}$

	y_1	y_2	y_3
1	0	0	0
2	0	1	1
3	1	1	1
4	1	1	0
5	1	0	0

The number of internal variables needed is equal to the minimum number of τ-partitions which cover all the $\pi = \overline{q_i, q_{i+1}; q_j}$. The following theorem gives the minimum number of partitions where $\lceil x \rceil$ equals the smallest integer greater than or equal to x.

Theorem 2 - The number of τ-partitions needed to cover the π-partitions for an n state time slot decoder is $\lceil \log_2 n \rceil$ if $n \neq 2^s - 1$ and $\lceil \log_2 n \rceil + 1$ if $n = 2^s - 1$ (s = 1, 2 ...).

Proof - The following set of τ-partitions satisfy the theorem for n = 4,5,6,8.

n = 4

$\tau_1 = \overline{q_1, q_2}; \overline{q_3, q_4}$

$\tau_2 = \overline{q_1, q_4}; \overline{q_2, q_3}$

n = 5

$\tau_1 = \overline{q_1, q_2}; \overline{q_3, q_4, q_5}$

$\tau_2 = \overline{q_1, q_5}; \overline{q_2, q_3, q_4}$

$\tau_3 = \overline{q_1, q_4, q_5}; \overline{q_2, q_3}$

n = 6

$\tau_1 = \overline{q_1, q_2}; \overline{q_3, q_4, q_5, q_6}$

$\tau_2 = \overline{q_1, q_2, q_3, q_6}; \overline{q_4, q_5}$

$\tau_3 = \overline{q_1, q_5, q_6}; \overline{q_2, q_3, q_4}$

n = 8

$\tau_1 = \overline{q_1, q_2, q_3, q_4}; \overline{q_5, q_6, q_7, q_8}$

$\tau_2 = \overline{q_1, q_4, q_5, q_8}; \overline{q_2, q_3, q_6, q_7}$

$\tau_3 = \overline{q_1, q_2, q_7, q_8}; \overline{q_3, q_4, q_5, q_6}$

For each given set of τ-partitions $q_1 = q_n(\tau_i)$ if and only if $i \neq 1$.

Now it will be shown if there exist sets of τ-partitions for n states ($2^{s-1} \leq n \leq 2^s$ and $n \neq 2^s - 1$) such that the number of τ-partitions is $\lceil \log_2 n \rceil$ and $q_1 = q_n(\tau)$ for all but one of the τ-partitions, then there exists a set of τ-partitions for m states ($2^s + 1 \leq m \leq 2^{s+1}$ and $m \neq 2^{s+1} - 1$) such that the number of τ-partitions is $\lceil \log_2 m \rceil = s + 1$ and $q_1 = q_m(\tau)$ for all but one of these new τ-partitions. Since the τ-partitions are given with the stated properties for s = 3, then the proof will proceed by induction on s.

Let m be any number such that $2^s + 1 \leq m \leq 2^{s+1}$ and $m \neq 2^{2+1} - 1$. Then it is a simple matter to show there exist integers n_1 and n_2 such that $2^{s-1} \leq n_1, n_2 \leq 2^s$; $n_1, n_2 \neq 2^s - 1$; and $n_1 + n_2 = m$. Let $\tau_1^1, \tau_2^1, \tau_s^1$ and $\tau_1^2, \tau_2^2, \ldots, \tau_s^2$ be the set of τ-partitions for a time slot decoder of n_1 and n_2 states respectively. Assume $q_1 = q_{n_p}(\tau_k^p)$ for k = 2, 3, --, s (p = 1, 2) and $q_1 \neq q_{n_p}(\tau_1^p)$ for p = 1, 2. Add n_1 to the subscripts of all the states in the τ-partitions $\tau_1^2, \tau_2^2, --, \tau_s^2$

to obtain the set of τ-partitions $\{\tau_1^3, \tau_2^3, \text{-- } \tau_s^3\}$. (Note these partitions will form valid partitions for n_2 states with the states being relabeled). Construct the s + 1 τ-partitions for m states in the following manner. For i = 1, 2, ---, s let

$$\tau_i = \{\overline{q | q = q_{n_1}(\tau_i^1) \text{ or } q = q_{n_1+1}(\tau_i^3)};$$

$$\overline{q | q \neq q_{n1}(\tau_i^1) \text{ or } q \neq q_{n_1+1}(\tau_i^3)}\}$$

If either n_1 or $n_2 = 2^{s-1}$, let $\tau_s^1 = \{\overline{q_1, q_2, \text{--- } q_{n_1}}\}$ or $\tau_s^3 = \{\overline{q_{n_1}\text{--}, q_m}\}$ respectively. Let

$$\tau_{s+1} = \{\overline{q_1, q_2, \text{-- } q_{n_1}}; \overline{q_{n_1+1}, q_{n_1+2}, \text{-- } q_{n_1+n_2} = q_m}\}$$

It is obvious $q_1 = q_m(\tau_i)$ for i = 1, 2, 3 -- s since $q_1 \neq q_{n_1}(\tau_1^1)$ and $q_1 \neq q_{n_2}(\tau_1^2)$; and $q_1 = q_{n_1}(\tau_i^1)$ and $q_1 = q_{n_2}(\tau_i^2)$ for i = 2, 3, ... s. Also it is obvious $q_1 \neq q_m(\tau_{s+1})$, which completes one part.

It now must be shown that these τ-partitions $\{\tau_1, \tau_2, \text{-- } \tau_{s+1}\}$ cover all π partitions of the type $\pi = \{\overline{q_i, q_{i+1}}; \overline{q_j}\}$. There are three different basic cases:

1. (a) $1 \leq i, i+1, j \leq n_1$ or
   (b) $n_1 + 1 \leq i, i+1, j \leq m$
2. (a) $1 \leq i, i+1 \leq n_1$ and $n_1 + 1 \leq j \leq m$ or
   (b) $1 \leq j \leq n_1$ and $n_1 + 1 \leq i, i+1 \leq m$
3. (a) $i = n_1$ and $i + 1 = n_1 + 1$ $(i \leq j \leq m)$ or
   (b) $i = m$ and $i + 1 = 1$ $(1 \leq j \leq m)$

Case 1(a) - Let τ_k^1 be a τ-partition such that $q_i = q_{i+1}(\tau_k^1)$ and $q_j \neq q_i(\tau_k^1)$ then $q_i = q_{i+1}(\tau_k)$ and $q_j \neq q_i(\tau_k)$ by definition of $\tau_k (1 \leq k \leq s)$. Hence τ_k covers π and similarly for 1(b).

Case 2(a) - Since $q_i = q_{i+1}(\tau_{s+1})$ and $q_j \neq q_i(\tau_{s+1})$, then τ_{s+1} covers π and similarly for 2(b).

Case 3(a) - By definition of τ_k(k = 1, 2, --- s) $q_{n_1} = q_{n_1+1}(\tau_k)$. Assume there exists no τ_k such that $q_j \neq q_{n_1}(\tau_k)$, which implies $q_j = q_{n_1}(\tau_k^1)$ or $q_j = q_{n_1+1}(\tau_k^3)$ for k = 1, 2, -- s. Either case is impossible since any two states must appear in different blocks of at least one τ-partition in the set for τ_k^1 of τ_k^2. Thus there exists a τ_k such that $q_{n_1} = q_{n_1+1}(\tau_k)$ and $q_j \neq q_{n_1}(\tau_k)$. Thus τ_k covers π and similarly for 3(b).

This proves the theorem except when $m = 2^s - 1$ for $s = 2,3,..$ By the first part of the proof there exist s τ-partitions $\{\tau_1, \tau_2, -- \tau_s\}$ for a machine of 2^s states. We now proceed to construct s + 1 partitions from the s partitions for a machine on 2^s states. Remove state q_{m+1} from each τ_k to form τ_k^1 for k = 1, 2, -- s. Let $\tau_{s+1}^1 = \{\overline{q_1, q_m}; \overline{q_2, q_3, -, q_{m-1}}\}$. We now determine if $\pi = \{\overline{q_i, q_{i+1}}; \overline{q_j}\}$ is covered by at least one of the τ_k^1. Two different cases exist.

1. $q_i \neq m$
2. $q_i = m$, $q_{i+1} = 1$

Case 1 - Let τ_k be the τ-partition which covers $\pi = \{\overline{q_i, q_{i+1}}; \overline{q_j}\}$ where i, i+1, j $\neq$ m + 1. Then τ_k^1 will obviously cover π for k = 1, 2, -- s.

Case 2 - τ_{s+1} will obviously cover the partition $\pi = \{\overline{q_m, q_1}; \overline{q_j}\}$. This proves the theorem.

It is quite easy to show when the number of states of the time slot decoder equals 2^s that any grey code when used to code the states will induce the necessary partitions to insure no function hazards exist.

ELIMINATION OF LOGIC HAZARDS

The following definitions will be used for elimination of hazards due to delays. Definition 5 is equivalent to the logic hazard definition of Eichelberger [1] or Howe and Coates [3].

Definition 5 - A network contains a static logic hazard for the input change A to B if and only if

(1) f(A) = f(B)
(2) No static function hazard exists for input change A to B.
(3) During the input change A to B, a spurious hazard pulse may be present on the output.

Definition 6 - A network contains a dynamic logic hazard for input change A to B if and only if

(1) $f(A) \neq f(B)$
(2) No dynamic function hazard exists for input change A to B.
(3) During the input change A to B a momentary "0" output and a momentary "1" output may appear.

Eichelberger [1] has shown that any sum-of-products realization of a function will be free of (static) logic hazards if and only if the realization contains all prime implicants. The way we have defined the output function Z_j^* such that it equals

1 only for internal state s_j it will consist of one prime implicant. Hence by Eichelberger Z_j^* will contain no static logic hazard. Now we must insure no dynamic logic hazard exists.

We assume m internal variables are needed. Then $Z_j^*(y_1, y_2, -- y_m) = y_1^* y_2^* -- y_m^*$ where y_i^* equals y_i or $\bar{y}_i$. Our function Z_j^* is realized with one "AND" gate.

Theorem 3 - $Z_j^*(y_1, y_2, -- y_m) = y_1^* y_2^* -- y_m^*$ will contain no dynamic logic hazards for any allowable transition when realized with one "AND" gate.

Proof - Assume Z_j^* does contain a dynamic logic hazard. Then for some internal state other then the final or initial state Z_j^* must be 1 for the "AND" gate to be momentarily 1. This is impossible since Z_j^* equals 1 for only one state s_j. Hence a contradiction.

SUMMARY

We have shown it is possible to code the states of a time slot decoder to eliminate all function hazards and usually no more than the minimum number of internal variables is needed. Also each output function when realized with one "AND" gate will contain no logic hazards. By using the procedure outlined in the proof of Theorem 2, it is possible to easily generate the final τ-partitions; and thus a state assignment for any time slot decoder.

The next step is the extension of the results presented here to the general case of any Moore model synchronous sequential circuit. This problem is somewhat more involved since a particular output (z) may contain an arbitrary number of ones (not just one) and elimination of static and dynamic hazards becomes more difficult.

1. E. B. Eichelberger, "Hazard Detection in Combinational and Sequential Switching Circuits," IEEE International Conf. Rec. on Switching Theory and Logical Design, Vol. 12, pp. 111-121, 1964.

2. J. H. Tracey, "The Internal State Assignment for Asynchronous Sequential Machines," IEEE Trans. Electronic Computers, Vol. EC-15, pp. 551-560, August, 1966.

3. A. B. Howe and C. L. Coates, "Logic Hazards in Threshold Networks," IEEE Trans. Computers, Vol. C-17, pp. 238-251, March, 1968.

THE USE OF CONNECTION GRAPHS FOR THE DETECTION OF DIGITAL FAULTS

DONALD R. SCHERTZ
Department of Electrical Engineering and Technology
Bradley University
Peoria, Illinois

GERNOT A. METZE
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois

ABSTRACT

Faults in combinational digital systems can be represented by a set of Boolean equations based on classes of indistinguishable faults. These equations are interpreted by means of connection graphs which are used for the generation of efficient test sets.

INTRODUCTION

The detection of failures in combinational digital systems has been the subject of a number of investigations. (1)(2)(3) These studies have been directed primarily toward the development of algorithms for the derivation of testing procedures for specific networks. In this paper, the approach is somewhat more theoretical in nature and leads to the presentation of heuristic techniques for the generation of "good" sets of diagnostic tests with a minimum amount of computation.

A system failure may be thought of as a transformation of the desired system into some other system. If the faulty system has a response which differs from that of the fault-free system under at least one set of input conditions, then the failure is said to be detectable. If the transformed systems resulting from two distinct failures have different responses for at least one set of input conditions, then the two failures are said to be distinguishable.

Fault Diagnosis of digital systems may be thought of as a process of partitioning the set of all possible failure transformations of a given system. Two distinct cases are of interest; fault detection which refers to a partition that separates the failure-free system from all others, and fault location which refers to a partition that separates those pairs of systems resulting from failures in separate modules.

In this paper, the fault model used to represent circuit failures will be a "line stuck at 1, line stuck at 0" model. If a line is stuck at 1 (sa1), it is considered to be permanently tied to the logical 1 level. Similarly, a line stuck at 0 (sa0) is considered to be permanently tied to the logical 0 level. It is further assumed that either of the faults

(sal or sa0) can occur at any gate input or output or at a primary input or output of the circuit. It is henceforth assumed that any physical failure in a system has been represented in terms of the appropriate fault model. Clearly, some physical failures such as shorts between lines cannot be represented by this model; however, it has been shown that all normal logic gate failures can be accurately modeled in this manner. (4) A shorthand notation is used for the fault model; the fault "line i sal" is written i-1, and the fault "line i sa0" is written i-0.

FAULT COLLAPSING

Fault collapsing is the process of deriving implication relationships between the detection of various system faults. (5) For example, the following implication relationships apply to an AND gate:

any input sa0 $\Leftrightarrow$ output sa0

any input sal $\Rightarrow$ output sal.

These relationships, which are readily specified for any combinational circuit, form a partial ordering relative to the detection of faults. In this manner, classes of indistinguishable faults are found before deriving dignostic tests rather than after. The faults for the circuit of Figure 1a are shown in Figure 1b with those in the same class linked together. In this figure, each pair of faults is drawn in the same relative position as the corresponding line in the circuit diagram. The upper fault in each case represents the line sal. This graphical representation of the faults is referred to as the fault class structure.

The fault classes resulting from any circuit are tree-like in form. Consider the segment of a circuit covered by a single class; one or more faults correspond to lines which carry information into the segment and a single line corresponds to the line which carries information out of the segment. These faults are referred to respectively as initial and final faults for the class. (In a class consisting of a single fault, that fault is both initial and final.) A class which has two or more initial faults is always formed from the bidirectional implication relationship between the input and output faults for a combinational gate. Such a class is said to span the gate.

Neighboring fault classes can be situated so that the final fault in one class corresponds to the same line as one of the faults in another class. In this case, the two classes are said to be strongly connected.

If two classes are situated so that a chain of strongly connected classes exists between them, they are said to be connected. In Figure 1b, there are two connected sets of classes; one set consisting of classes 3 and 4, and the other set consisting of the remaining fault classes.

The tree-like characteristic of fault classes carries over to connected sets of classes. Progressing through a connected set in the forward direction, paths may converge but will never diverge. This is another way of saying that the portion of the original circuit which is represented by a connected set contains no fanout point. The connected set representation thus subdivides a circuit into fanout-free or tree-like segments.

Since the connected set corresponds to a fanout-free segment of the circuit, a single line carries information out of the circuit segment. If the connected set contains more than two classes, then one of the classes having its final fault on this line spans a gate. The other class on this line is the final class in the connected set. The initial classes for a connected set are defined in an analogous manner. If the connected set consists of only two classes, then either may be considered as the final class. (The remaining class is then considered to be initial.)

FAULT CLASS REPRESENTATION

The total state of a combinational circuit in the presence of faults could be fully specified by giving the logic level carried on each line in the circuit. The theoretical evaluation of the state of any line under fault conditions would have to include the effects of the faults corresponding to that line. Unfortunately, the three-valued nature of the condition of a line (normal, sa1, sa0) makes such an evaluation relatively complicated if the effects of several faults are to be considered simultaneously. (3)

This difficulty in the fault representation is resolved by the use of fault classes rather than lines as coordinates for the representation of the total state of a circuit. The line value associated with the final fault in a class is used to represent that class. The line on which the final fault is located may carry the logic value corresponding to that fault either as a result of the occurrence of a fault in the class or as a result of conditions on other lines which affect the logic value of the line considered.

The total state of the circuit is represented by the binary state vector $\{S\}$ where $S_i=1$ whenever the line on which the final fault of class i is located carries the value corresponding to that fault. Similarly, the fault condition in the circuit is represented by a binary fault vector $\{F\}$ where $F_i=1$ whenever a fault in class i occurs. The fault functions are defined so that they model the actual circuit failures considered. The state functions, however, are also partially determined by the signals applied to the circuit inputs. Thus S_i may be 1 either because a fault in class i occurred ($F_i=1$) or because other conditions in the circuit cause the line to take on the representative value. It is obvious from consideration of the direction of information flow in the original circuit that S_i depends on conditions in the part of the circuit which lies between class i and the primary inputs and therefore may be specified in terms of the fault function for class i and the state functions for the classes which directly precede class i.

This technique of state representation is probably best developed by means of a simple example. The circuit of Figure 1a which realizes the function $Z = W\,\overline{X} + X\,\overline{Y}$ has the fault class structure of Figure 1b. The final fault class of the connected set closest to the output is class 10 which contains the single fault j-0. The state function corresponding to class 10 must thus represent the condition "line j carries the logic value 0." This condition can hold either because fault j-0 occurs or because the line is given that value by conditions in the remainder of the circuit. The logic value 0 on line j is the complement of the logic value 1 on line j, and the logic value 1 on line j is the condition for the state function corresponding to class 9. Therefore, the state function for class 10 can be written as

$$S_{10} = F_{10} + \overline{S_9}.$$

As was stated above, the state function for class 9 represents the condition "line j carries the logic value 1." This can occur either as a result of a fault in class 9 or as a result of conditions in the circuit between class 9 and the primary inputs. Consideration of the original circuit shows that line j will carry the logic value 1 whenever either line h or line i carries the logic value 0. The latter conditions are the complements of the conditions for the state functions corresponding

to classes 7 and 8, hence

$$S_9 = F_9 + \overline{S_7} + \overline{S_8}.$$

In a similar manner, the state functions for classes 7 and 8 are found to be

$$S_8 = F_8 + \overline{S_2} + \overline{S_6}$$

$$S_7 = F_7 + \overline{S_1} + \overline{S_5}.$$

Classes 5 and 6 represent faults on destination lines from a fanout point. The state function for class 6 represents the condition "line e carries the logic value 1". Line e will carry the logic value 1 whenever line b carries the logic value 1. However, the state function for class 4 represents the condition "line b carries the logic value 0". The appropriate definition for the state function for class 6 becomes

$$S_6 = F_6 + \overline{S_4}.$$

Similarly, the state function for class 5 is written

$$S_5 = F_5 + S_4.$$

Class 4 may be treated in a manner analogous to that of class 10. The state function in this case is

$$S_4 = F_4 + \overline{S_3}.$$

The three remaining classes are initial classes which lie on primary input lines. Class 3 represents the condition "line b carries the logic value 1". This condition may be caused by a fault in class 3 or by the input condition "input X is 1". Thus, the state function for class 3 is

$$S_3 = F_3 + X.$$

Similarly, the state functions for classes 1 and 2 are

$$S_2 = F_2 + \overline{Y}$$

$$S_1 = F_1 + W.$$

The output function Z may be related to the state function for the final class of the connected set closest to the output by noting that the state function corresponds to the logic value 0 on line j. Therefore, the desired functional relationship is

$$Z = \overline{S_{10}}.$$

The rules for the evaluation of state functions may be summarized quite simply. The state function for any class is the disjunction of the fault function for that class with an expression involving the immediate predecessors of the class. This expression takes one of three forms depending on the location of the class whose state function is being evaluated.

1) For any class except an initial class in a connected set, the expression is the disjunction of the complements of the state functions for all classes strongly connected to the given class at its initial faults.

2) For an initial class which lies on a destination line from a fanout point, the expression is either the state function for the final class of the connected set preceding the fanout point or its complement. If the initial fault of the initial class and the final fault of the final class represent lines stuck at the same logic value, the state function itself is used. If the faults represent lines stuck at opposite logic values, the complement of the state function is used.

3) For an initial class which lies on a primary input line, the expression is either the input proposition or its complement. If the initial fault of the initial class represents a line sa1, the input proposition itself is used. If the fault represents a line sa0, the complement of the input proposition is used.

These three cases cover all fault structures normally encountered. A special case arises in circuits which contain succeeding gates of the same logic type (AND feeding AND, OR feeding NOR, etc.). The treatment of this special fault class structure may be found in a previous publication. (5)

An expression for the circuit output in terms of the fault functions and primary inputs may be found by starting with the output and successively substituting the definition of each state function. For our example, this expression is

$$Z = \overline{F_{10}}\,(\,F_9 + \overline{F_7}\,((\,F_1 + W\,)\,(\,F_5 + F_4 + \overline{F_3}\,\overline{X}\,)) \\ \overline{F_8}\,((\,F_2 + \overline{Y}\,)\,(\,F_6 + \overline{F_4}\,(\,F_3 + X\,)))).$$

Thus the behavior of the circuit under all fault conditions is described by a Boolean equation rather than an equation involving three-valued functions such as that found by Pogue. (3) This equation may be manipulated to generate tests for faults or to simulate the effect of faults on circuit behavior. The fact that it is a Boolean equation rather than a three-valued representation means that all the tools of Boolean Algebra may be utilized in this procedure.

CONNECTION GRAPH INTERPRETATION

Any Boolean function may be represented as a connection graph which is isomorphic to a contact network that realizes the given function. For each appearance of a literal in the function to be represented, the graph has one edge whose weight is that literal. In the case of the fault expression for the circuit output, one edge is present for each occurrence of a circuit input and one edge is present for each occurrence of a fault function. An edge forms a connection when its weight is 1 and does not form a connection when its weight is 0. The value of the output function is 1 whenever a path is formed between the two terminal nodes of the graph and is 0 whenever no path exists between the terminal nodes.

The dual graph with corresponding edge weights complemented represents the complemented output function. The two graphs may be used in conjunction for the derivation of diagnostic tests. If the circuit has more than one output, then there is a pair of graphs corresponding to each output.

A fault is detected by a test if the circuit output in the presence of the fault differs from that of the fault-free circuit. Every test completes a path between the terminal nodes in one of the two output graphs. When a test is applied to a fault-free circuit, all edges corresponding to complemented fault functions are opened. The edges corresponding to circuit inputs are opened or shorted in accordance with the value of the test input. Any path that is formed is thus made up of edges corresponding to complemented fault functions and to input variables. Any edge whose weight is 0 does not form a connection. This is equivalent to removing that edge from the representation of the graph.

If a single path is completed, then the removal of any edge in that path will separate the two terminal nodes. An edge having weight $\overline{F_j}$ is removed if a fault in class j occurs. However, the occurrence of this fault may add or remove more than one edge (the literal F_j may appear more than once in the output function). A test which completes a path between terminal nodes thus detects a fault in class j if and only if the removal of all edges corresponding to $\overline{F_j}$ and the simultaneous addition of all edges corresponding to F_j break the path between terminal nodes. If multiple paths are completed between nodes, then every such path must be broken if the test is to detect the fault.

The condition for the detection of a fault may thus be expressed in terms of the set of paths between terminal nodes which are completed by a

test. A fault in class j will be detected by a test if and only if the removal of all $\overline{F_j}$ edges breaks every path formed by the test and the simultaneous addition of all F_j edges does not form any new paths connecting terminal nodes. For multiple faults, one need only consider the simultaneous addition and removal of edges corresponding to each class involved.

The application of connection graphs to fault detection will be illustrated with the aid of the graphs of Figure 2. These graphs represent the circuit of Figure 1. It should be noted that each graph contains a single edge corresponding to each fault class in the connected set nearest the output but contains multiple edges corresponding to those fault classes which precede a fanout point.

The test input (1,1,0) is a true vector for the circuit considered. The test completes a single path consisting of edges $\overline{F_{10}}$, $\overline{F_8}$, Y, $\overline{F_4}$, and X through the graph for Z = 1 and separates the terminal nodes in the graph for Z = 0. If a fault in classes 4, 8, or 10 occurs, the path will be broken. The literals F_8 and F_{10} appear only once in the graph for Z = 1 so the test detects faults in these two classes; that is, faults c-1, e-0, g-0, i-0, and j-0. However, the graph for Z = 1 contains an edge corresponding to the uncomplemented fault function F_4, and it is seen that the addition of this edge completes a second path thus inhibiting the detection of faults in class 4. It should be noted that the test input (0,1,0) would inhibit the completion of this secondary path and would thus detect faults in classes 4, 8, and 10.

It is easily verified that the test input (0,1,1) which completes a multiple path through the graph for Z = 0 detects faults in classes 2 and 9. The test input (1,1,1) completes a single path. This test detects every fault detected by (0,1,1) and in addition detects faults in classes 4 and 5.

The discussion of the preceding paragraphs illustrates the two conditions for the detection of a fault, i.e. the fault must break a path between terminal nodes, and no new path must be completed. These two conditions allow the following statements to be made concerning tests for the detection of single faults. As a primary objective, the test should form a single complete path between terminal nodes of the graph. If multiple paths are formed by the test, the section which is multiple should be as short as possible. (In some graphs representing redundant circuits it is impossible to complete a single path.) As a secondary objective, as few

off path edges as possible should be completed by the test. If multiple fault detection is of primary importance, these two objectives should be interchanged.

These arguments have been applied successfully to the manual selection of an efficient set of tests, for circuits having on the order of 20 lines. A programmed procedure using these techniques is clearly feasible.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grant NSF GK -1663.

REFERENCES

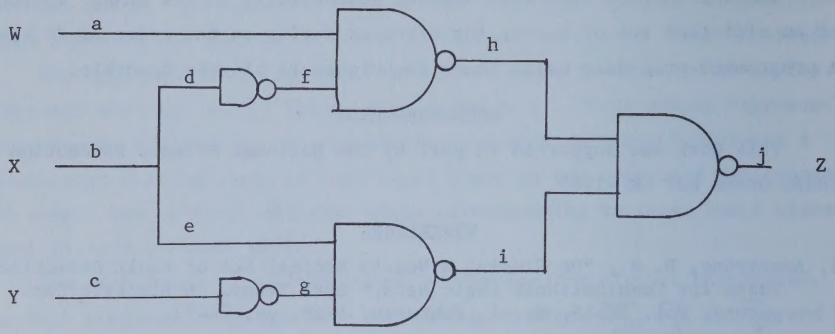
1. Armstrong, D. B., "On Finding a Nearly Minimal Set of Fault Detection Tests for Combinational Logic Nets," IEEE Trans. on Electric Computers, Vol. EC-15, No. 1, February, 1966, pp. 66-73.

2. Roth, J. P., Bouricius, W. G., and Schneider, P. R., "Programmed Algorithms to Compute Tests to Detect and Distinguish Between Failures in Logic Circuits," IEEE Trans. on Electronic Computers, Vol. EC-16, No. 5, October, 1965, pp. 567-580.

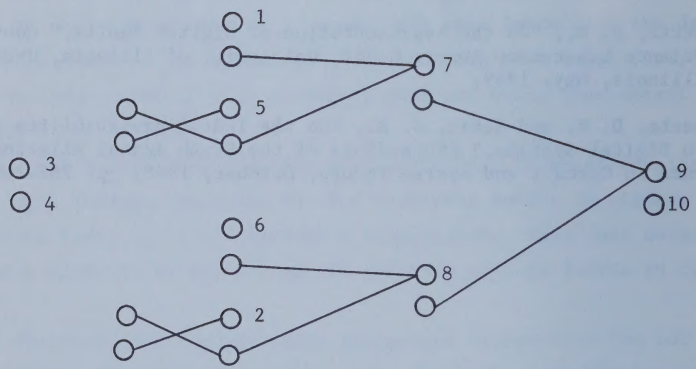
3. Poage, J. F., "Derivation of Optimum Tests to Detect Faults in Combinational Circuits," Mathematical Theory of Automata, Polytechnic Press, Brooklyn, N. Y., 1963, 483-528.

4. Schertz, D. R., "On the Representation of Digital Faults," Coordinated Science Laboratory Report R-418, University of Illinois, Urbana, Illinois, May, 1969.

5. Schertz, D. R. and Metze, G. A., "On the Indistinguishability of Faults in Digital Systems," Proceedings of the Sixth Annaul Allerton Conference on Circuit and System Theory, October, 1968, pp. 752-760.

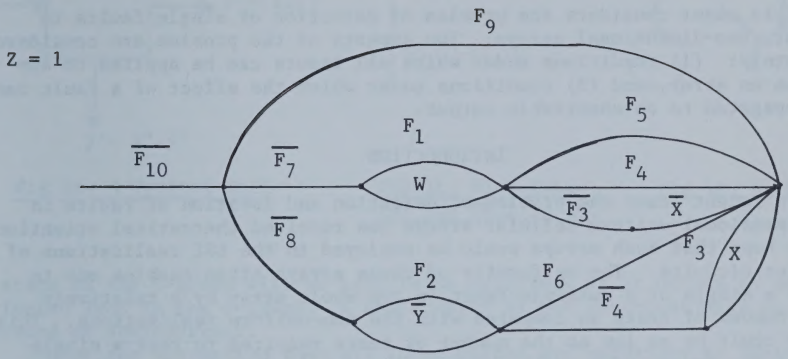


A.

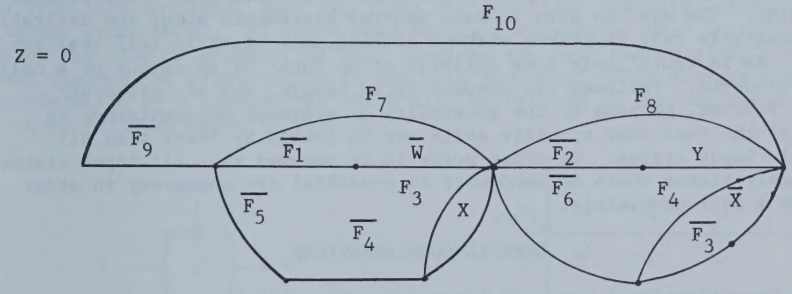


B.

Figure 1


Z = 1
F9
F10
F7
F1
W
F5
F4
F3
X
F8
F2
Y
F6
F4
F3
X

A.


Z = 0
F10
F9
F7
F1
W
F8
F2
Y
F5
F3
F4
X
F6
F4
X
F3

B.

Figure 2

FAULT DIAGNOSIS OF COMBINATIONAL CELLULAR ARRAYS

SHARAD CHANDRA SETH
Coordinated Science Laboratory
Urbana, Illinois

ABSTRACT

This paper considers the problem of detection of single faults in regular, two-dimensional arrays. Two aspects of the problem are considered separately: (1) conditions under which all inputs can be applied to any cell in an array, and (2) conditions under which the effect of a fault can be propagated to an observable output.

INTRODUCTION

In recent years the problem of detection and location of faults in two-dimensional uniform cellular arrays has received theoretical attention in the hope that such arrays would be employed in the LSI realizations of computer circuits. The uniformity of these arrays often enables one to detect a single or a multiple fault in the whole array by a relatively small number of tests as compared with the non-uniform realizations. This number could be as low as the number of tests required to test a single cell in the array, or a small multiple thereof.[1] Very little work has been done on the problem of location of a fault in an array, perhaps because it is considered economical to replace a whole faulty chip than to try to locate the fault(s) in it. The detection problem in an array can of course be approached by means of methods developed for non-uniform combinational circuits.[2,3,4] However such an approach is unpractical for the typical LSI circuits (with hundreds of gates and lines) so alternative methods must be found which exploit the regularity of the structure which is known to be there a priori.

In this paper the detection problem is considered from a theoretical viewpoint. The aim has been to make general statements about the desirable or undesirable cell functions without looking into specific cell realizations. As in Kautz' only line failures at an input or an output of a cell are considered. Failures are assumed to be single, and of "stuck-at" type. Further, because of the generality of approach no cognizance is taken of the fact that specific cells may be tested by fewer than all possible input states. In other words it is assumed that all input states (or simply states where no ambiguity is possible) are necessary in order to test a cell completely.

1. GENERAL CONSIDERATIONS

A two-dimensional array of identical cells will be considered. Each cell is assumed to have a horizontal and a vertical input and output. Except for the cells on the boundary each cell in the array feeds its horizontal output to the cell on its right and its vertical output to the cell below it. The external inputs are applied to the left and the upper terminal edges and the outputs are observed at the right and the lower terminal edges. A typical cell in the array is shown in Fig. 1a. The cell can be described by means of a "flow-table" as in Fig. 1b, or equivalently, by means of a graph (denoted as a G-graph in the subsequent discussion) as in Fig. 1c.

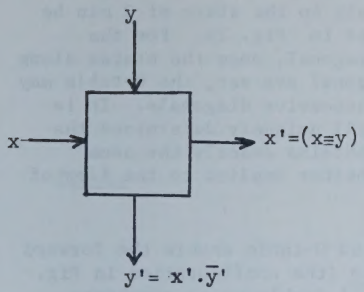


fig 1a. A Typical Cell

x \ y	0	1
0	(1,1)	(0,0)
1	(0,0)	(1,0)

fig 1b. Flow Table for the Cell in 1a.

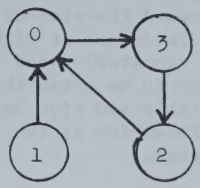


fig 1c. G-Graph

A state in the G-graph will be said to be recurrent if it occurs in an oriented loop in the G-graph. A state which is not recurrent will be denoted as nonrecurrent.

Under the assumption that all input states are necessary to test a cell, there are two conditions that must be satisfied to detect a fault in an array. These are:

(1) Every input state must be applicable to every cell in the array; and

(2) A sensitized path must exist from the fault location to one of the outputs.

The second condition makes it possible to detect faults from the boundary alone. Together these conditions are sufficient to detect single faults.

The signals in an array may be considered to flow from top to bottom and from left to right, that is, along the resultant direction of the main diagonal drawn NW-SE. It would seem natural, therefore, that in order to follow the flow of signals one must consider the cell states along the other diagonal at right angles to the main diagonal. This prompts the consideration of a corner-trio which will be defined as the configuration of three cells as shown in Fig. 2a.

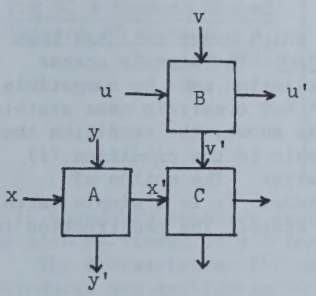


fig 2a. The Corner-Trio

xy \ uv	0	1	2	3
0	3	2	2	2
1	1	0	0	0
22	1	0	0	0
3	3	2	2	2

fig 2b. D-Table for Cell in 1a.

It will be noted that in fig. 2a, if the states of the cells A and B are fixed then the state of the cell C is automatically fixed. The mapping from the pair of states of A and B respectively to the state of C can be represented by means of a D-table as indicated in fig. 2b. For the forward transfer of signals along the main diagonal, once the states along an interface at right angles to the main diagonal are set, the D-table may be repeatedly used to set the states along successive diagonals. It is also to be noted that the flow-table for a cell uniquely determines the D-table and vice versa so that the D-table contains exactly the same information as the flow-table but in a form better applied to the flow of signals.

Just as the corner-trio and its associated D-table enable the forward propagation of signal, the inverse corner-trio (the configuration in Fig. 3a) and its associated D^{-1}-table (Fig. 3b) will enable us to propagate signals backwards, that is, in the direction opposite to the forward direction of signal flow considered above.

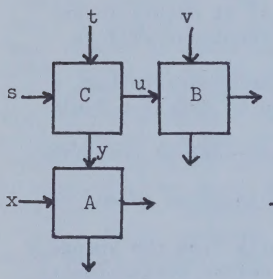


fig 3a. The Inverse Corner-Trio

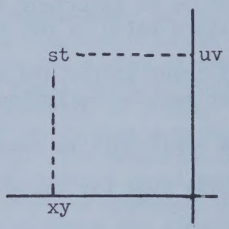


fig 3b. D^{-1}-Table

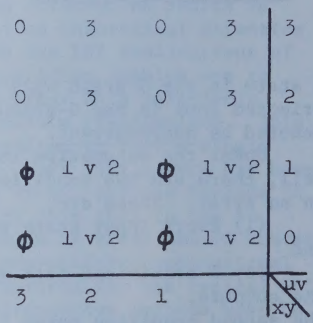


fig 3c. D^{-1}-Table for Cell in fig 1a.

The D^{-1}-table differs from the D-table in one essential aspect even though both of them represent essentially identical information. The difference arises because, unlike the mapping from the cell input to the cell output, the reverse mapping may not be one-one. Thus, in contrast with the D-table, the D^{-1}-table may contain empty entries as well as multiple entries (see Fig. 3c where the empty entries are denoted by ϕ and the multiple entries by the union of cell states).

A doubly infinite arrangement of cells in which every cell has been marked by its state will be called a tessellation. The signals across the horizontal (vertical) interfaces in a tessellation must be compatible from one row (column) to the next and this fact may constrain some state(s) from appearing in a tessellation. As Kautz[1] has shown, the condition that every state appear in a tessellation is isomorphic to the condition (i) above for detection of faults in an arbitrary array. The notion of recurrence of a state plays an important role here. It is a simple matter to construct a tessellation for each recurrent state. The construction is illustrated in Fig. 4 for the cell of Fig. 1a.

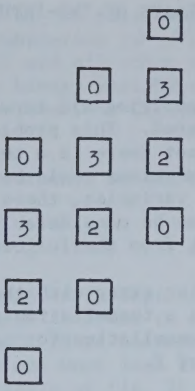


fig 4.

Thus a recurrent state can always be applied as inputs to any cell in an array. Unfortunately, the converse is not necessarily true for a non-recurrent state. It may or may not occur in a tessellation. The condition (i) reduces now to the condition on the existence of a tessellation for each one of the nonrecurrent states. It is conjectured that in general this may not be a solvable problem.[1]

A tessellation may consist of a finite recurring pattern. Such a pattern has been defined as a tessera.[5] Of particular interest is the diamond shaped tessera (hereafter called the D-tessera) of Fig. 5. The order of the D-tessera of Fig. 5 will be denoted as mxn

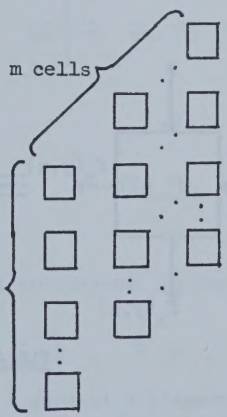


fig 5. A Diamond Shaped Tessera of order m x n

It is enough to restrict attention to only D-tessera because any tessera may also be viewed as a D-tessera of a suitable order.

The discussion so far has been general enough in that all the notions introduced are applicable to cells with any number of horizontal and vertical inputs. In the next section these notions are applied to two-input cells.

2. ARRAYS OF TWO-INPUT CELLS

EXISTENCE OF TESSELLATIONS

First the problem of applying all input states to any two-input cell in an array will be considered. This problem is equivalent to the problem of determining whether or not there is a tessellation for each state. Since each of the output functions could be any of the sixteen possible functions of the two input variables, there are 256 different cells to be considered. This number can be considerably reduced by considering duality and symmetry, as is evident from the following two theorems:

Theorem 1: The cellular array with the cell realizing the function-pair $(f_x(x,y), f_y(x,y))$ has a tessellation for each input state iff $(f_x^d(x,y), f_y^d(x,y))$ has a tessellation for each input state, where f^d denotes the dual function of f.

Proof: It is easy to see that if the cell-functions are duals then the tessellations are also duals, that is, with 1's and 0's interchanged. Therefore the tessellation for the state (x.y) for a cell yields a tessellation for the state $(\bar{x}.\bar{y})$ in the dual cell.

Theorem 2: The cellular array with the cell realizing the function-pair $(f_x(x,y), f_y(x,y))$ has a tessellation for each input state iff the array with the cell function-pair $(f_y(y,x), f_x(y,x))$ has a tessellation for each input state.

Proof: As is evident from Figs. 6a and 6b, the use of one cell function-pair instead of the other as stated in the theorem merely corresponds to an interchange of the two axes. Hence the theorem is obviously true.

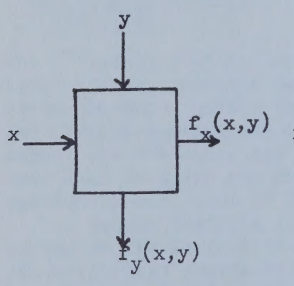


fig 6a.

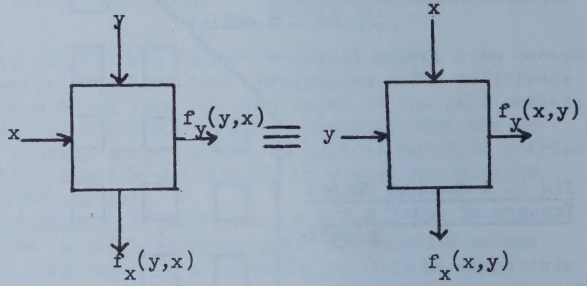


fig 6b.

Each of these theorems by itself allows a reduction of about a half in the number of function-pairs that need be considered. However, together the reduction is considerably less than a fourth because of the overlap between the two cases.

Further, there are certain classes of function-pairs about which it is possible to make a general statement regarding the existence of tessellations. Let a nodal function be defined as one which contains either a single zero or a single one in its truth-table representation. Consider

the function-pairs in which one of the pair, say f_x, is a nodal function. Then there is only one input combination (x^*, y^*) which is mapped to the constant value c (either 0 or 1) and all other combinations are mapped to $\bar{c}$ under f_x. Both x^* and y^* are binary variables and can assume either of the two values, c and $\bar{c}$. Let us consider the case under which $x^* = c$ and $x^* = \bar{c}$ separately.

Case 1: $x^* = c$. This corresponds to the situation where f_x is any of the following functions:

$$x+\bar{y} \quad , \quad x \cdot \bar{y} \quad , \quad x+y \quad , \quad \text{and} \quad x \cdot y.$$

In terms of the G-graph the situation is depicted in fig. 7a wherein the states have been shown in clusters in order to clearly indicate the constraints imposed. The state $(c\cdot\bar{y}^*)$ is nonrecurrent no matter how the constraints are satisfied hence one must look for a tessellation for this state. The partially filled D-table of fig. 7b may be used to see what would happen in a cell state $(c\cdot\bar{y}^*)$ is assumed in an array. This is shown in fig. 7c from which it is clear that the state $(c\cdot\bar{y}^*)$ can not be repeated horizontally to its left. Hence there can not be a finite tessera for the function-pairs: $(x+\bar{y}, -)$, $(x\cdot\bar{y}, -)$, $(x+y, -)$, $(x\cdot y, -)$.

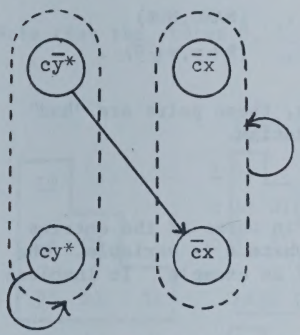


fig 7a.

	cy*	$c\bar{y}^*$	$\bar{c}x$	$\bar{c}\,\bar{x}$
cy*	c-	cx	c-	c-
$c\bar{y}^*$	$\bar{c}$-	$\bar{c}x$	$\bar{c}$-	$\bar{c}$-
$\bar{c}x$	$\bar{c}$-	$\bar{c}x$	$\bar{c}$-	$\bar{c}$-
$\bar{c}\,\bar{x}$	$\bar{c}$-	$\bar{c}x$	$\bar{c}$-	$\bar{c}$-

fig 7b.

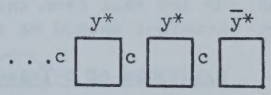


fig 7c.

Case 2: $x^* = c$. This corresponds to the situation where f_x is any of the following functions:

$$\bar{x}\cdot\bar{y} \quad , \quad \bar{x}+\bar{y} \quad , \quad \bar{x}\cdot y \quad , \quad \text{and} \quad \bar{x}+y.$$

It is not possible to make a general statement about this case without imposition of further constraints. Thus attention will be confined to a subcase shown in fig. 8a. The state (cy*) is nonrecurrent no matter how the G-graph of fig. 8a is completed hence one must look for a tessellation for this state. Fig. 8b shows the partially filled D-table for this case. Let the state (cy*) be assumed to occur in a tessellation. Fig. 8c shows how this assumption leads to a contradiction, hence the subcase corresponds to function-pairs without a tessellation for at least one non recurrent state.

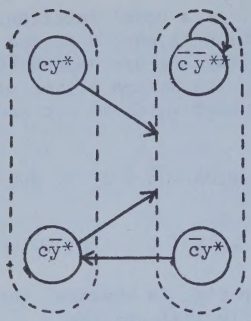


fig 8a.

	$\bar{c}y^*$	$c\bar{y}^*$	$\bar{c}x$	$c\bar{x}$
$\bar{c}y^*$	$\bar{c}y^*$	$c\bar{y}^*$	c-	c-
$c\bar{y}^*$	$c\bar{y}^*$	$\bar{c}y^*$	$\bar{c}$-	$\bar{c}$-
$\bar{c}x$	$\bar{c}y^*$	$c\bar{y}^*$	$\bar{c}$-	$\bar{c}$-
$c\bar{x}$	$\bar{c}y^*$	$c\bar{y}^*$	$\bar{c}$-	$\bar{c}$-

fig 8b.

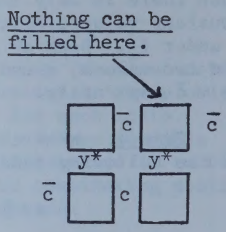


fig 8c.

There are sixteen function-pairs corresponding to this subcase. These are:

(NOR,$\bar{x}$) ,	(NOR,1) ,	(NOR,$\bar{x}+y$) ,	(NOR,NAND)
($\bar{x}$y,x) ,	($\bar{x}$y,0) ,	($\bar{x}$y,xy) ,	($\bar{x}$y,x$\bar{y}$)
(NAND,$\bar{x}$) ,	(NAND,0) ,	(NAND,$\bar{x}$y) ,	(NAND,NOR)
($\bar{x}+y$,x) ,	($\bar{x}+y$,1) ,	($\bar{x}+y$,x+y) ,	($\bar{x}+y$,x+$\bar{y}$) .

One can say that from the viewpoint of diagnosis, these pairs are "bad" and therefore should be avoided in the logical design.

Existence of D-Tessera

2Xn D-Tesserae: An algorithm can be given in terms of the entries of the D-table to determine all 2Xn D-tesserae where n is variable. The algorithm can best be described with the help of an example. It involves the following three steps in sequence:

(i) Form unordered pairs from the (i,j)th and the (j,i)th entries of the D-table (fig. 10a).

(ii) Pass 1: Cross off all the identical pairs.

Pass j(j > 1): If the entry (i,j) was crossed off during pass j-i then cross off all occurrences of either (i,j) or (j,i) in the table.

The successive passes for our example are shown in fig. 10b.

(iii) If step (ii) results in crossing off of all the entries in the table then the nonrecurrent state(s) in the G-graph do not have a 2Xn D-tessera. If there are still uncrossed entries left then construct a testing-graph in which these entries are represented as vertices. The testing-graph is constructed in the same way as the G-graph was constructed from the flow-table. Only the states that occur as one of the pair corresponding to a recurrent vertex in the testing graph, have a 2Xn D-tessera. Fig. 10c shows the testing graph for the example. It also illustrates how the 2Xn tessera may be constructed using the information from the testing graph.

The proof that the algorithm gives all the 2Xn D-tesserae is rather

obvious once the steps are properly understood.

Example: Let $x' = x \equiv y$ and $y' = x + \bar{y}$

	0	1
0	1,1	0,0
1	0,1	1,1

fig 9a. Flow Table for the example

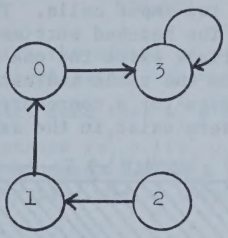


fig 9b. G-Graph

	0	1	2	3
0	3	2	3	3
1	1	0	1	1
2	1	0	1	1
3	3	2	3	3

fig 9c. D-Table

Note that the states 0, 1, and 2 are nonrecurrent.

1	12		
2	13	01	
3	33	21	31
	0	1	2

fig 10a.

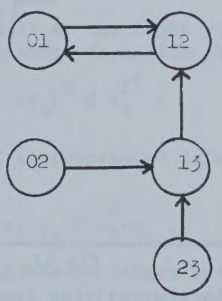
PASS I

1	12		
2	13	01	
3	X	21	31

PASS II

1	12		
2	13	01	
3	X	21	31

fig 10b.



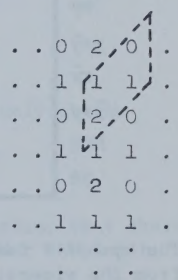


fig 10c.

Higher Order D-tesserae: In principle one can find all possible D-tesserae of order mxn by considering all possible m-tuples of states and drawing an oriented graph in which these are represented by vertices. An arrow is drawn from the vertex $(x_1,x_2 \ldots x_m)$ to the vertex $(y_1,y_2 \ldots y_m)$ if and only if the diagonal interface of the input states, $(x_m x_1 x_2 \ldots x_m)$ results in an interface $(y_1,y_2 \ldots y_m)$ in the forward direction. Then all the mxn (n variable) D-tesserae are represented by loops in this graph and every loop of length n corresponds to a D-tessera of order mxn. This method, when applied in practice, soon becomes impracticable as the number m goes up because of the great number of m-tuples involved in the computation.

The D^{-1}-table can often be used in eliminating a number of m-tuples

to be considered. If the back-drive of signals from an m-tuple fails in all possible cases that m-tuple (plus all the additional m-tuple that may occur during a back-drive) can be eliminated from consideration. A judicious use of D- and D^{-1}-tables leads to efficient searching procedure for higher order D-tesserae.

The methods described in this section were applied to complete the 16 X 16 function-pair table for two-input cells. The results are displayed in the table shown in fig. 11. The hatched entries in the table indicate the pairs for which no tessellations exist for one or more of the non-recurrent states. The numbers in the table indicate the minimum number m in the order, mxn, of the D-tessera for a nonrecurrent state. The symbol ∞ indicates that no finite tessera exist in the said case.

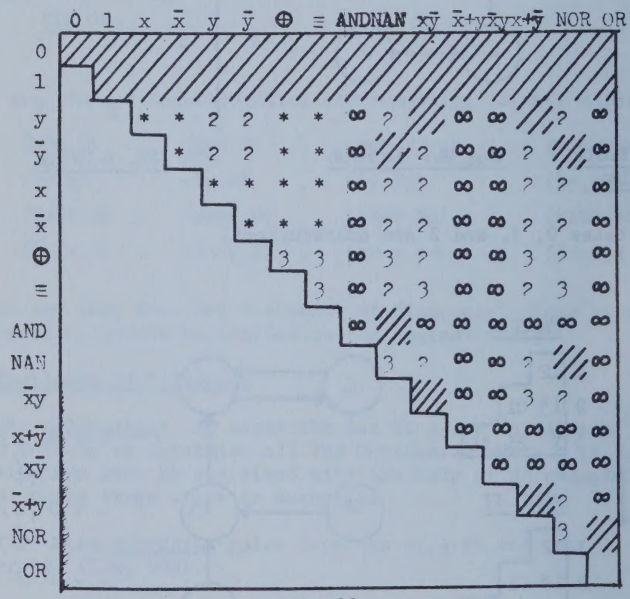


fig 11.

The symbol * indicates that all the states in the G-graph are recurrent. From the viewpoint of diagnosis, the function-pairs corresponding to the hatched entries are obviously undesirable. When the order of the D-tessera is small, it is possible to apply all possible states to all the cells in an array by a small number of tests by successively shifting the array in a tessellation. Thus, the pairs corresponding to ∞ entry in the table would, in general, require more number of tests. One can therefore say that the function-pairs with f_x = x·y, or x+y, or x·ȳ, or x+ȳ OR fy = x·y, or x+y, or x̄y, or x̄+y should be avoided as cell functions.

FAULT PROPAGATION

Attention will be confined to unidirectional (horizontal, or vertical) propagation of single faults. In other words, it is enough to consider only cascades of cell. Let the array be assumed to have a recurring pattern in the direction of fault propagation, in the absence of a fault. Consider a horizontal cascade of cells as follows:

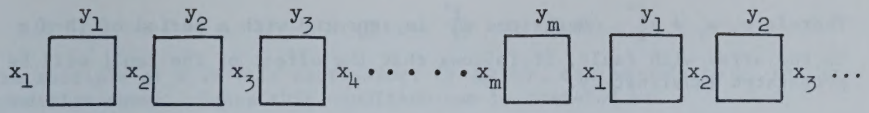


fig 12.

The conditions under which the effect of a horizontal fault is propagated indefinitely will be investigated. Assume that the cell (x_1,y_1) changes to (x_1^f,y_1) under a fault, such that $x_1^f \neq x_1$. Let the cascade change correspondingly to the states $(x_1^f,y_1)(x_2^f\ y_2)\ \cdots\ (x_m^f\ y_m)$ etc. The following notation will be used to denote the action of the primary inputs $x_1,y_1,y_2,\ldots,y_{j-1}$ to determine the x_j output:

$$x_j = x_1 \cdot (y_1 y_2 \cdots y_{j-1}), \quad j > 1 .$$

Since the horizontal alphabet is finite, there must exist two integers, M and N, N > M, such that

$$x_1^f \cdot (y_1 y_2 \ldots y_m)^N = x_1^f \cdot (y_1 y_2 \ldots y_m)^M = x_1^{f'} \quad \text{(say)} .$$

For convenience let the smallest pair (M,N) be chosen such that M+N is minimum. Then the following theorem can be stated.

Theorem 3: The effect of a fault at the input of the cell (x_1,y_1) is propagated indefinitely in the horizontal direction iff

$$x_1 \cdot (y_1 y_2 \ldots y_m)^M \neq x_1^{f'} .$$

Proof: (a) Suppose the effect is propagated indefinitely in the horizontal direction. Then,

$$\text{for all } i, \quad x_i \cdot y_i \neq x_i^f \cdot y_i .$$

Since the horizontal input alphabet is finite this implies that there must exist two integers, M and N, such that the signals under the fault form a recurring pattern with a period of (N-M) m, and

$$x_1 \cdot (y_1 y_2 \ldots y_m)^M \neq x_1^{f'}$$

where $x_1^{f'}$ is defined as before,

(b) Next suppose that

$$x_1 \cdot (y_1 y_2 \ldots y_m)^M \neq x_1^f (y_1 y_2 \ldots y_m)^M .$$

Since the pattern in the horizontal direction is assumed to recur in the absence of a fault, for all integers k the following must be true:

$$x_1 \cdot (y_1 y_2 \ldots y_m)^k = x_1 .$$

Therefore, $x_1 \neq x_1^{f'}$. But since $x_1^{f'}$ is repeated with a period of (N-M)m in the array with fault, it follows that the effect of the fault will be propagated indefinitely.

Theorem 4: In a single rail cascade with periodic input pattern, the linear functions, involving x(y) input nontrivially, are both necessary and sufficient for the horizontal (vertical) propagation of the effect of a fault.

Proof: The sufficiency is obvious because by definition the linear function is complemented whenever one of its arguments is complemented. To prove the necessity, consider a cascade of two cells as shown in fig. 13.

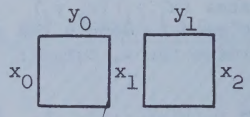


fig 13.

The transformation of the x input realized by the cell under the y_o input can be represented by a logic matrix (see Preparata[6]) L_o and that under y_1 input can be represented by the logic matrix L_1, where

$$L_o = \begin{bmatrix} 1 & 0 \\ a_1 & b_1 \end{bmatrix}, \quad \text{and } L_1 = \begin{bmatrix} 1 & 0 \\ a_2 & b_2 \end{bmatrix}.$$

The combined transformation can be written as

$$\begin{bmatrix} 1 & 0 \\ a_2 & b_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ a_1 & b_1 \end{bmatrix} \begin{bmatrix} 1 \\ x_0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_2 \end{bmatrix}.$$

Thus $x_2 = (a_2 \oplus a_1 b_2) \oplus b_1 b_2 x_o$; in general $x_n = F_o \oplus F_1 x_o$ where F_o is independent of x_o and $F_1 = b_1 b_2 \ldots b_n$. For propagation x_n must depend on x_o, therefore a necessary condition for propagation is that

$$b_1 b_2 \ldots b_n = 1, \quad \text{that is,} \quad \underline{b_j = 1 \quad \text{for all } j.}$$

Let us determine the cell functions implied by this condition. If y is considered a variable instead of a parameter, the logic matrix for the cell functions can also be written as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ a & b & c & d \\ - & - & - & - \\ - & - & - & - \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \\ xy \end{bmatrix} = \begin{bmatrix} 1 \\ x' \\ y' \\ x'y' \end{bmatrix}$$

where the - entries are unspecified. From the expansion of the logic matrix it is found that

$$x' = (a \oplus cy) \oplus x(b \oplus dy) \ .$$

The multiple of x in this expression, $(b \oplus dy)$, corresponds to b_j in the condition above. Hence this condition can be restated as

$$\underline{b \oplus dy = 1} \ .$$

This further implies that either (b = 0 and dy = 1) or (b = 1 and dy = 0). The first of these imposes a value on y and can be disregarded because we have assumed that all values of y are necessary to test a cell. The second condition, that is, (b = 1, and dy = 0) can be satisfied by (b = 1 and d = 0). This corresponds to the linear functions stated in the theorem, as can be easily verified.

ACKNOWLEDGMENT

I wish to thank Professor G. Metze and Professor F. Preparata for discussion of various ideas during the period of this work.

REFERENCES

1. Kautz, W. H., "Testing Faults in Combinational Cellular Logic Arrays," Proceedings of the Eighth Annual Symposium on Switching and Automata Theory, Austin, October 1967, pp. 161-174.

2. Roth, J. P., "Diagnosis of Automata Failures: A Calculus and a Method," IBM Journal, July 1966, pp. 278-291.

3. Poage, J., "Derivation of Optimal Tests to Detect Faults in Combinational Circuits," Proceedings of the Symposium on Mathematical Theory of Automata, Brooklyn Polytechnic, April 1962, pp. 483-528.

4. Kautz, W. H., "Fault Testing and Diagnosis in Combinational Digital Circuits," IEEE Transactions on Computers, Vol. C-17, No. 4, pp. 352-366, April 1968.

5. Tammarru, E., "Testing of Combinational Logic Cells in Large-Scale Arrays," Submitted to the IEEE Computer Group Repository, March 19, 1969.

6. Preparata, F. P., "State-Logic Relations for Autonomans Sequential Networks," IEEE Transaction on Electronic Computers, Vol. EC-13, October 1964, pp. 542-548.

IMPLEMENTATION AND EXTENSION OF MULTI-DIMENSIONAL PATH SENSITIZING IN A SIMULATION AND DIAGNOSIS SYSTEM

S.A. SZYGENDA and G.C. GOLDBOGEN
Computer Science Department
University of Missouri-Rolla, Rolla, Missouri

ABSTRACT

This paper considers the problem of diagnostic test generation for large combinational logic nets. The foundations of multi-dimentional path sensitizing, as well as some new implementations and optimization techniques, are considered. Results are given for various test cases which were used for system evaluation.

I. PRELIMINARY

The system considered in this paper has as its primary objective the generation of diagnostic tests for large combinational logic nets. Combinational test generation has been considered extensively elsewhere[1,2,5]. However, many of the techniques proposed do not possess the capability of generating tests for all non-redundant faults, or are not easily implemented for large nets.

The implementation to be discussed is part of a total simulation and diagnostic test generation system[4,5]. As such, it will be required to perform the following functions:

1) Generate a test for any specified non-redundant fault occuring in a large combinational logic net. Since the combinational test generator will comprise only a portion of the total system, efficiency and speed are considered to be of the utmost importance.
2) Generate tests that can distinguish between two or more given faults.
3) Possess the capability of verifying arbitrary input or output specifications. This may involve considering a specified subset of inputs and/or outputs and then generating complete input-output vectors that are consistent with the original specifications.
4) The combinational test generation algorithm should be usable as a heuristic for sequential test generation.

The present version of the program is restricted to considering single logical faults of the stuck-at-one (s-a-1) or stuck-at-zero (s-a-0) type. This limitation can be easily modified to consider other classes of faults, if necessary.

To accomplish the first function, a multi-dimensional path sensitizing technique has beem implemented[3]. This technique utilizes the d-algorithm, proposed by Roth[3], with modifications to achieve program efficiency and flexibility. The algorithm has the ability of generating a test, if one exists, for any fault in the combinational logic. A brief discussion of the general character of this algorithm will ensue a more detailed discussion of the implementation will be given in Section II.

Assume that a fault, F, exists in the circuit. It is desired to generate a test, T, capable of detecting, at least, the fault under consideration. Then it is necessary to

propagate the fault conditions to primary outputs. This is usually referred to as the forward-trace phase of the procedure.

On the other hand, it is usually necessary to have the test specified in terms of the primary inputs to the circuit. To satisfy this condition it will be necessary to "back-up" the specified output values to determine the critical primary input values. This process is called the back-trace phase of the algorithm.

After a test has been generated, using the test generation algorithm, it is simulated to obtain whatever additional information is inherent in the test. The results of this "ping-pong" strategy between test generation and simulation is monitored by a decision module of the system. This module commands the two sub-systems as needed and keeps track of tasks completed and remaining. The operation of the system, to satisfy the conditions considered above, can best be illustrated by a simple example.

Consider a circuit, C, containing faults M_1, M_2, M_3,..., M_n. The failure free circuit will be labeled M_0. The objective is to generate a sequence of tests capable of distinguishing the circuits containing faults from M_0. The process consists of the following steps:

a) The decision module assumes control and monitors the stack of faults. When the fault list is empty this function is complete.
b) If the fault list is not empty, the top entry in the stack (M_i) is considered.
c) The combinational test generator is called upon to generate a test for the specified fault.
d) The generated test (T_j) is simulated to determine which other faults may be detectable by the test. These faults are then removed from the fault list.
e) The generated test is added to the test sequence with a pointer to the list of faults detectable by the test. This information is used for partitioning and fault catalog production.

This process is demonstrated by the following diagram:

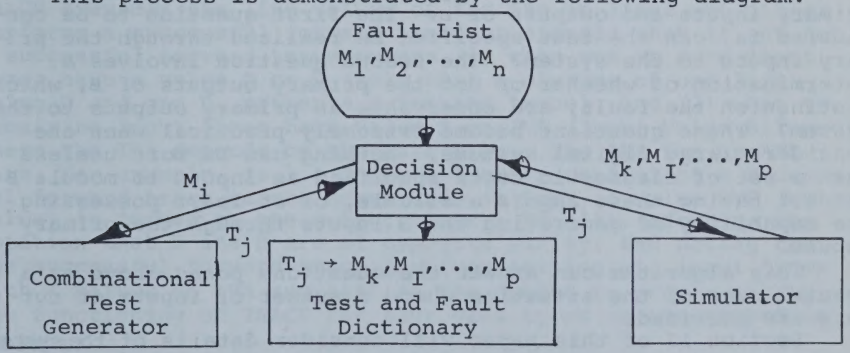


Figure 1. Combinational Test Generation

Before considering the second function, it is noteworthy that the system is implemented in such a manner that the performance of the various functions differs only in the initialization step of the process. The second function extends the

path sensitizing technique. This extension provides the ability of generating a test (if a test exists) capable of distinguishing between two or more faults. This algorithm utilizes the physical relation of the faults and the various connecting paths between the faults, and/or the propagation of the effects of the faults. When a path is chosen, a starting point is determined and the multi-dimensional path sensitizing technique is used to complete the test.

The next function involves compatability rather than test generation. Consider the possibility that a partial list of primary inputs or outputs are specified for the combinational circuit. This can occur if certain inputs are constrained to a logical zero or one value to conform to overall system operational requirements. The algorithm will take these subsets of values and generate a complete input vector that is compatible with the specified inputs and outputs.

The following illustration will demonstrate the usefulness of this function.

Consider the system described in Figure 2, with faults in module B being considered.

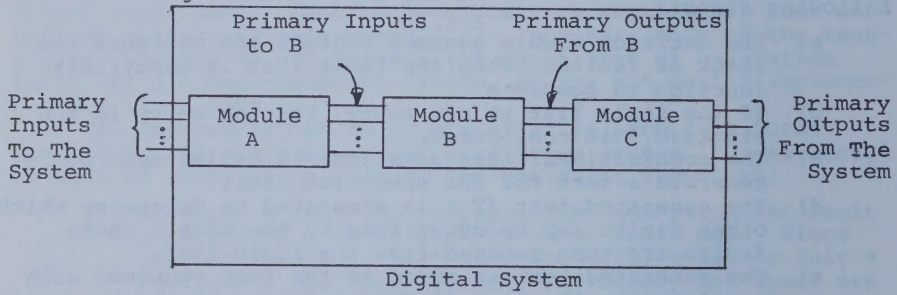


Figure 2. Compatability of Primary Inputs and Outputs

The test generated for a given fault will be in terms of primary inputs and outputs of B. The first question to be considered is, can the test specified be realized through the primary inputs to the system? The second question involves a determination of whether or not the primary outputs of B, which distinguish the fault, are observable as primary outputs to the system? These questions become extremely practical when one considers large digital systems. Nothing can be more useless than a set of diagnostic tests specified as inputs to module B and not having these inputs available, or at least possessing the capability of generating the B inputs through the primary inputs.

This algorithm can answer the questions posed above as a special case of the situation where a subset of inputs or outputs are provided.

Section II of this paper will consider details of the system implementation to perform the functions considered thus far. This will be followed by a discussion of the results and evaluations of the algorithms.

One required function of the test generation system remains to be considered. It is not grouped with the other functions because it is controlled by the overall simulation and diagnosis system and not the combinational test generation segment. This

function utilizes the combinational test generator as an aid for sequential logic test generation. The combinational algorithm can be utilized for this function in at least two ways. First, we can utilize the following heuristic: break the feed-back loops in a sequential circuit, consider the broken loops to represent primary inputs and outputs to the circuit, and then generate a test for a fault in the corresponding combinational net. After the test is generated, the feedback loops are reconnected and the test is applied by the simulator. For many cases the simulation will show that this particular test does not detect the fault. However, for sequential logic with limited feedback, the process is more successful. In lieu of any general algorithm for sequential test generation, which can be implemented for large nets, this approach appears to be acceptable.

The second possibility, presently under investigation, is the extension of the combinational testing procedures to specific classes of sequential circuits; for example, circuits containing only set-reset flip-flops. Preliminary results in this area appear to offer some encouragement.

II. IMPLEMENTATION

Consider a combinational circuit, C, consisting of N gates and L levels. The gates in the first level are input gates, that is, primary inputs to C, and the gates in level L are output gates, that is, primary outputs of C. Seven gate types are presently considered: AND, OR, NAND, NOR, NOT, INPUT, or OUTPUT, but the method presented is easily extended to other gate types. First, the gates are numbered sequentially by levels. The line (or lines) forming the output of gate I is called line I. A line will have one of the following line values: X, undetermined line value; 0, constant value 0; 1, constant value 1; D, variable value, 1 for the fault free circuit, 0 for the faulty free circuit, $\bar{D}$, variable value, 0 for the fault free circuit, 1 for the faulty circuit.

The forward trace is accomplished by the TRACF subroutine which accepts a gate number, gate type, fan in, and fan out. TRACF examines the line values of the fan in lines and either performs a successful forward trace or signals that it can not. A successful forward trace occurs (1) when one or more input lines have a value D or $\bar{D}$ (note that a mixture of one line being D and one $\bar{D}$, prevents a forward trace), and (2) when the remaining input lines have value X or IVA (Input Value Allowed) where the IVA depends on the gate type. The IVA for an AND gate, for example, is 1. Under all other conditions a successful forward trace cannot occur and, in fact, a value OVF (output value) is forced on the output lines. Finally, adopting the notation that D and $\bar{D}$ are of opposite parity, and noting that the successful forward trace involves setting all input lines with X values to IVA and setting the output to proper parity, the functioning of TRACF for four gate types is summarized in Table I.

The back trace function is accomplished by the TRACB subroutine, which accepts a gate number, gate type, fan in, and fan out. An attempted back trace thru a gate has one of four outcomes: (1) successful back trace, only one possible way to accomplish back trace, (2) successful back trace, K ways to accomplish back trace, where K is the number of input lines with value x, (3) back trace already completed, (4) back trace

impossible because of a conflict. The functioning of TRACB can be summarized in the following table II showing the outcomes as numbered above under various input line conditions:

Gate Type	IVA	OVF	Successful Output
AND	1	0	Same parity as input
OR	0	1	Same parity as input
NAND	1	1	Opposite parity from input
NOR	0	0	Opposite parity from input

Table I. TRACF Function

Gate Type	AND		OR		NAND		NOR	
Value to be Traced Back	1	0	1	0	1	0	1	0
All K Lines = X	(1)	(2)	(2)	(1)	(2)	(1)	(1)	(2)
Some=IVA, Remaining K=X	(1)	(2)	(2)	(1)	(2)	(1)	(1)	(2)
Some=$\overline{\text{IVA}}$, Remaining K=X	(4)	(3)	(3)	(4)	(3)	(4)	(4)	(3)
Some=IVA, Some=$\overline{\text{IVA}}$	(4)	(3)	(3)	(4)	(3)	(4)	(4)	(3)
Any 2 = D and $\overline{\text{D}}$	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)

Table II. TRACB Function

Note: for (1) Set K lines to IVA

(2) Set 1 of K lines to $\overline{\text{IVA}}$

Consider the following circuit with fault 1 (F_1) present (but not F_2):

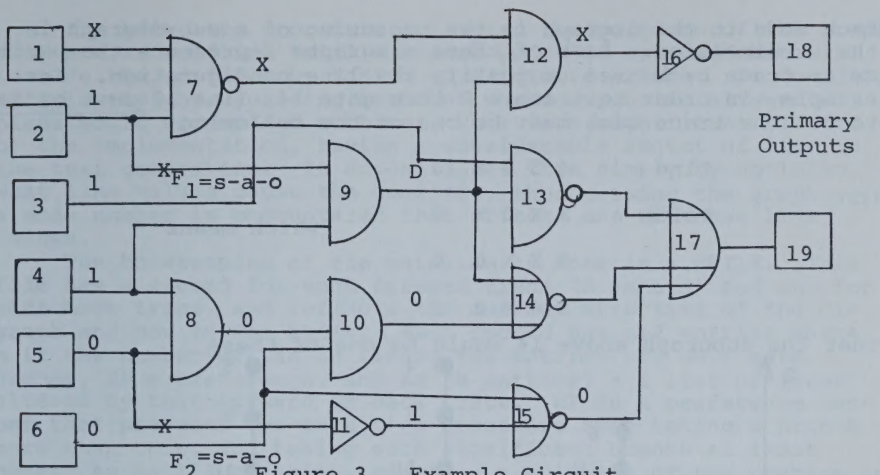


Figure 3 Example Circuit

It is desired to generate a test for fault F_1. In this circuit, there are many possible combinations of forward and back traces, some of which lead to a solution. In order to consider all of these combinations systematically and efficiently, the method used is to dynamically build and search a three dimensional tree or digraph described below.

The values of line 9 may be traced forward thru one of 3 single paths {9,12,16,18}, {9,13,17,19} or {9,14,17,19}, or one of 4 multiple paths which are combinations of the single paths 2 at a time, and 3 at a time. The forward trace portion of the tree corresponding to these paths is in the X Y plane. Thus, initially the value D is thru gate 9 which has fan out 12,13,14. This corresponds to the tree.

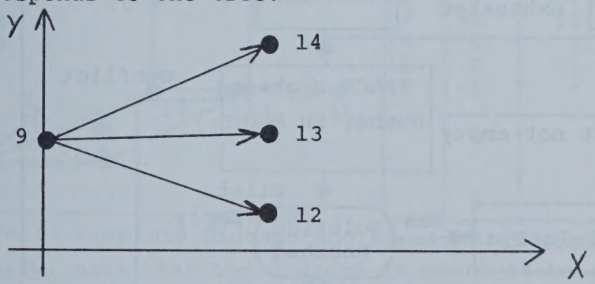


Figure 4 Digraph Partial Base

This digraph may grow and modify itself by adding and deleting branches until a solution is reached. Thus, the portion of the digraph for the solution in the X Y plane will be a subset of:

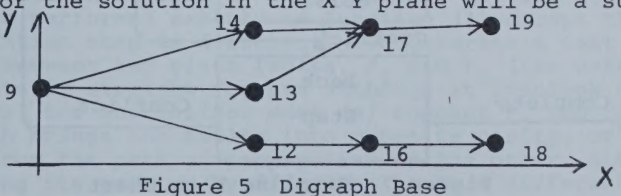


Figure 5 Digraph Base

Each node in the digraph is the beginning of a sub-digraph in the z dimension. Each of these subgraphs represents the particular trace back used to justify the line configuration. For example, in order to trace a D thru gate 14, line 10 must be set to 0. The trace back must be one of the following:

Line	4	5	6	8	10	
Value	X	X	0	X	0	which means
	0	X	X	0	0	
	X	0	X	0	0	

that the subgraph above 14 would be one of these

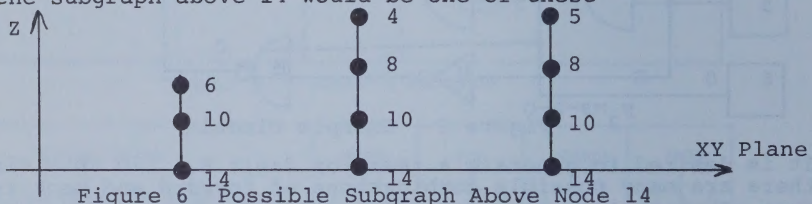


Figure 6 Possible Subgraph Above Node 14

The complete digraph is constructed node by node, back tracing at each node. The leading points of the digraph are entries in the ELIST of the program. The leading points of the subgraph being constructed in the z dimension are entries in the BLIST. The process of the mainline program is flow charted below:

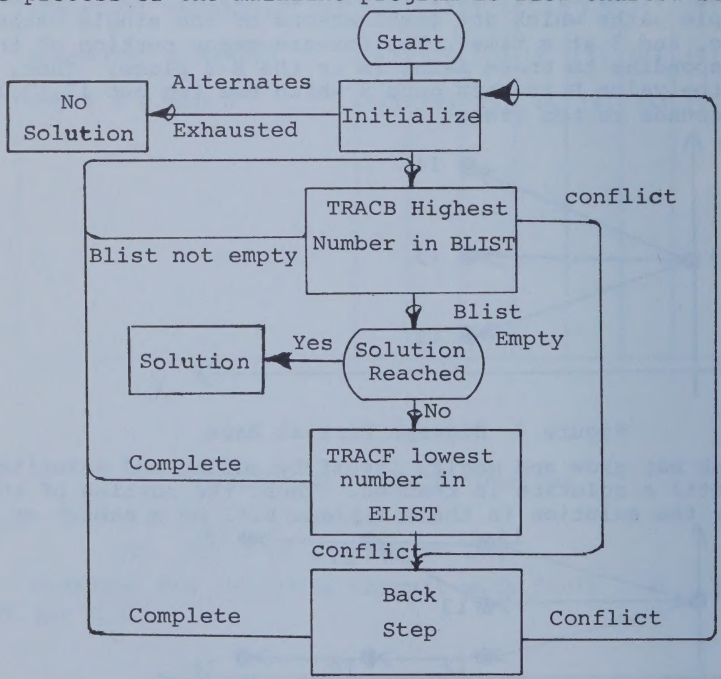
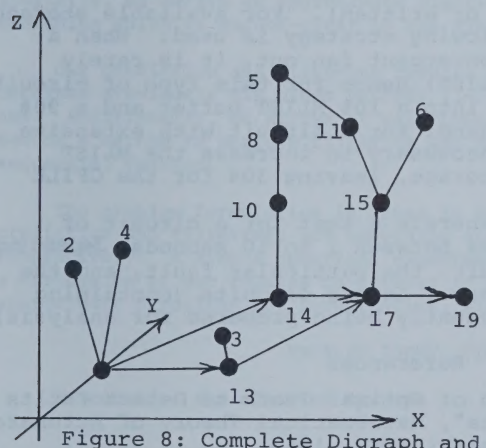


Figure 7 Mainline Flow Chart

The back step procedure collapses the portions of the graph that are in conflict until a branch point is reached that will resolve the conflict by taking an alternate branch. It does so, efficiently, by bypassing branches where exploring alternate plans would accomplish nothing. This is a significant feature of the implementation, saving a considerable amount of time in the test generation. It accomplishes this simply by noticing what line values cause the conflict, then erasing the graph until a node number is encountered that effects one of these line values.

The bookkeeping of the mainline is done in a MFILE. This file has a record for each forward trace (E record) and one for each back trace, and reflects the current structure of the digraph and how it was built. Each record has m+2 entries where m is the number of fan in lines; the entries are NG = gate number, NP = preference, and LN (M entries) = a list of lines altered by that forward or back trace. NP is a preference number that prevents the back step procedure from taking a branch more than once, and taking each significant branch at least once. As an illustration, consider a solution of the problem in our example circuit, the corresponding digraph, and the MFILE given below:

SOLUTION: line	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
values	X	1	1	1	0	0	X	0	D	0	1	X	$\overline{D}$	$\overline{D}$	0	X	$\overline{D}$	X	$\overline{D}$



MFILE

NG	NP	LN
9	-1	2,4
12	0	0,0
13	-2	3
14	-2	10
10	1	8
8	0	5
17	-1	15
15	0	6,11
11	0	0
19	-1	0

Figure 8: Complete Digraph and Final MFILE for Solution

Finally, note that the digraph is meant to be an aid in visualizing the process and that the mainline works with the MFILE, CFILE (which describes the circuit), BLIST and ELIST only. The "back step" process is done by back stepping up the MFILE to the appropriate record.

As noted in section I, the functions (2) and (3) of the system are performed exactly as function (1) except that the initialization step is different. To generate a test to distinguish between two given faults, F_1 and F_2 (the extension to many faults is straight forward although it involves some combinatorics) the initializer must (a) connect F_1 and F_2 by a path which brings the faults into opposite parity, or (b) cut off one from the path of propagation of the other, either by terminating its path or by directing it to a different primary output.

Thus, considering F_1 and F_2, in the example, the initializer may choose to cut off F_2 by setting line 6 to zero or it may propagate both to line 17 where they meet in opposite parity.

The final function, the setting of the lines to the specified subsets of primary inputs and/or primary outputs is an initialization step; the mainline very simply will verify the specified values if the configuration is valid. Space consideration prevent a more extensive explanation of these functions.

III. CONCLUSION AND RESULTS

Fortran IV was utilized for the programs described above. The storage requirements are as follows. The mainline program initializer, TRACF, and TRACB each require approximately 1K words of core. The array requirements vary with the circuit size. If N is the number of gates in the circuit, and M is the maximum fan in or fan out, then the CFILE is less than Nx(2M) words while the MFILE never grows larger than M records M+2 words per record = N x (M+2) words.

Minimum core requirements (overlaying the subroutines and the initializer) are slightly in excess of 2K words (this includes small buffers).

Of course, if enough storage is available, it is desirable to have both files in core. At the other extreme, a minimal buffer size would be enough for one record for each file (containing the last record read or written). For available storage of intermediate size the following strategy is used: When a circuit has little or no reconvergent fan out, it is rarely necessary to read from the MLIST; hence for this type of circuit, divide the available storage into a 10% MLIST buffer and a 90% CFILE buffer. On the other hand, for a circuit with extensive reconvergent fan out, it is necessary to increase the MLIST buffer to 70% of available storage, leaving 30% for the CFILE buffer.

The time necessary to generate a test for a circuit of approximately 100 gates varies between 1 to 10 seconds, depending on the structure of the circuit, the particular fault, and the availability of files in memory. Larger circuits (containing more than 1000 gates) are presently being prepared for analysis).

IV. References

1. Poage, J.F., "Derivation of Optimal Tests to Detect Faults in Combinational Circuits", Mathematical Theory of Automata, Polytechnic Press, Brooklyn, N.Y., 1963.

2. Armstrong, D.B., "On Finding a Nearly Minimal Set of Fault Detection Tests for Combinational Logic Nets", IEEE Trans. on Elec. Computers, Vol. EC-15, No. 1, 1966.

3. Roth, J.P., "Diagnosis of Automata Failures: A Calculus and A Method", IBM Journal of Research and Development, Vol. 10, 1966.

4. S.A. Szygenda, "A Software Diagnostic System for Test Generation and Simulation of Large Digital Systems", Proc. of the N.E.C. Annual Conference, December 1969.

5. S.A. Szygenda, and D. Rouse, "Functional Models for Simulation of Large Digital Systems", Proc. of the Joint Conf. on Math. and Computer Aids to Design, October, 1969.

SEQUENTIAL ESTIMATION FOR DISCRETE-TIME NONLINEAR SYSTEMS

J. S. MEDITCH
Boeing Scientific Research Laboratories
Seattle, Washington 98124

ABSTRACT

The problem of state and parameter estimation for noisy discrete-time nonlinear dynamic systems is examined from the viewpoint of marginal maximum likelihood estimation. Approximate algorithms for sequential prediction, filtering, and smoothing are developed. The former two are in agreement with previous results; the latter is new. A technique for iterative-sequential filtering and smoothing using Newton's method is indicated. A numerical example is included to illustrate the results.

INTRODUCTION

In this paper, we examine the problems of prediction, filtering, and smoothing for discrete-time nonlinear dynamic systems from the viewpoint of marginal maximum likelihood estimation. We obtain algorithms for sequential prediction and filtering by utilizing approximations which are first-order in the system dynamics and second-order in the measurement function. Similarly, we derive an approximate algorithm for reverse-time sequential smoothing which is second-order in the system dynamics.

Our prediction and filtering results are in agreement with those obtained previously via other methods [1-3]. Our smoothing results are new. Additionally, we arrive at relations which can be utilized along with Newton's method to perform filtering and smoothing in an iterative-sequential fashion.

The problem formulation is given in Section 2, and the algorithms for prediction and filtering, and smoothing are derived in Sections 3 and 4, respectively. A simple numerical experiment is presented in Section 5, and a discussion of our results is offered in Section 6 where we also give an approximate algorithm for forward-time, single-stage, sequential smoothing.

PROBLEM FORMULATION

We consider the class of systems

$$\begin{aligned} x_{i+1} &= f(x_i) + w_i \\ z_i &= h(x_i) + v_i \end{aligned} \tag{1}$$

for $i=0,1,\ldots$ where i is the discrete-time index; x is an n-vector, the state; w is an n-vector, the disturbance; z is an m-vector, the measurement; and v is an m-vector, the measurement error. In Eqs. (1) and (2), f and h are, respectively, n and m-dimensional, vector-valued, twice continuously differentiable functions of the indicated variables. It is understood that these two functions may also depend on known time-varying parameters although this explicit time-dependence has been omitted to simplify the notation.

We assume that $\{w_i, i=0,1,\ldots\}$ and $\{v_i, i=0,1,\ldots\}$ are independent,

zero mean, gaussian, random sequences with $E[w_j w_k'] = Q_k \delta_{jk}$, $E[v_j v_k'] = R_k \delta_{jk}$, $E[w_j v_k'] = 0$, for all discrete time-points $j,k \geq 0$. Here, E denotes the expected value, prime the transpose, and δ_{jk} the Kronecker delta. Further, Q_k and R_k are, respectively, $n \times n$ and $m \times m$ covariance matrices both of which are assumed to be positive definite.

We take the initial state x_0 to be a zero mean, gaussian, random vector which is independent of $\{w_i, i=0,1,\ldots\}$ and $\{v_i, i=0,1,\ldots\}$ and has a positive definite covariance matrix P_0.

Our problem consists in obtaining an estimate of x_k, $k \geq 0$, given the sequence of measurements $\{z_0,z_1,\ldots,z_j\}$ where $j \geq 0$. Specifically, denoting such estimates by $\hat{x}_{k|j}$, we shall examine the three cases of (1) *prediction*, where $j=k-1$ and $k=0,1,\ldots$; (2) *filtering* where $j=k$ and $k=0,1,\ldots$; and (3) *smoothing* where $j=N$ is fixed and $k=N-1,\ldots,0$; and obtain algorithms for computing these estimates sequentially in time.

We approach the problem from the viewpoint of marginal maximum likelihood estimation. In particular, we let $Z_j = \{z_0,z_1,\ldots,z_j\}$ and consider the likelihood functions

1. $L(x_k,Z_{k-1}) = \log p(x_k|Z_{k-1})$ for prediction,

2. $L(x_k,Z_k) = \log p(x_k|Z_k)$ for filtering,

and

3. $L(x_k,x_{k+1},Z_N) = \log p(x_k,x_{k+1}|Z_n)$ for smoothing.

In these relations, $p(\cdot|\cdot)$ is the marginal conditional probability density function and the logarithm is to the base e.

Our work here is based on consideration of only first and second-order moments, i.e., it is assumed that third and higher-order moments are negligible. This assumption is a common one in nonlinear estimation theory, but is difficult to justify in practice.

PREDICTION AND FILTERING

For a fixed $k \geq 1$, we have

$$\begin{aligned} p(x_k|Z_{k-1}) &= \int_{R^n} p(x_k,x_{k-1}|Z_{k-1})dx_{k-1} \\ &= \int_{R^n} p(x_k|x_{k-1},Z_{k-1})p(x_{k-1}|Z_{k-1})dx_{k-1} \\ &= \int_{R^n} p(x_k|x_{k-1})p(x_{k-1}|Z_{k-1})dx_{k-1} \end{aligned} \tag{3}$$

where the last line follows from the preceding one since the conditional random vector x_k given x_{k-1} is independent of Z_{k-1}. In Eq. (3), R^n denotes euclidean n-space and $dx_{k-1} = dx_{k-1}^1 \cdots dx_{k-1}^n$ where x_{k-1}^ℓ is the

ℓ^{th} component of x_{k-1}.

Since $\{w_i, i=0,1,\ldots\}$ is an independent, zero mean, gaussian, random sequence, it follows that

$$p(x_k|x_{k-1}) = \alpha_k \exp[-\tfrac{1}{2}||x_k - f(x_{k-1})||^2_{Q^{-1}_{k-1}}] \tag{4}$$

where

$$\alpha_k = 1/\sqrt{(2\pi)^n |Q_{k-1}|}$$

and $|\cdot|$ denotes the determinant, $[\ ]^{-1}$ the matrix inverse, and $||\xi||^2_A$ the quadratic form $\xi'A\xi$.

We now assume that $p(x_{k-1}|Z_{k-1})$ can be expanded in an Edgeworth series [1,4] in terms of its first two moments. We thus write

$$p(x_{k-1}|Z_{k-1}) \tilde{=} \beta_{k-1} \exp[-\tfrac{1}{2}||x_{k-1} - \hat{x}_{k-1|k-1}||^2_{P^{-1}_{k-1|k-1}}] \tag{5}$$

where

$$\beta_{k-1} = 1/\sqrt{(2\pi)^n |P_{k-1|k-1}|}\ .$$

Here, $\hat{x}_{k-1|k-1}$ is the conditional mean and $P_{k-1|k-1}$ the conditional covariance matrix.

Higher-order moments can be included in Eq. (5) if desired, but our development is based on first and second moments only.

Utilizing the first-order approximation

$$f(x_{k-1}) \tilde{=} f(\hat{x}_{k-1|k-1}) + f_x(\hat{x}_{k-1|k-1})(x_{k-1} - \hat{x}_{k-1|k-1}), \tag{6}$$

where $f_x(\hat{x}_{k-1|k-1})$ is the $n \times n$ Jacobian matrix of $f(x_{k-1})$ evaluated at $x_{k-1} = \hat{x}_{k-1|k-1}$, in Eq. (4), and substituting the result along with Eq. (5) into Eq. (3), we obtain after some lengthy, but straightforward manipulation, the expression

$$p(x_k|Z_{k-1}) = \gamma_k \exp[-\tfrac{1}{2}||x_k - f(\hat{x}_{k-1|k-1})||^2_{P^{-1}_{k|k-1}}] \tag{7}$$

where

$$\gamma_k = 1/\sqrt{(2\pi)^n |P_{k|k-1}|}$$

and

$$P_{k|k-1} = f_x(\hat{x}_{k-1|k-1}) P_{k-1|k-1} f'_x(\hat{x}_{k-1|k-1}) + Q_{k-1}. \tag{8}$$

Trivially, the marginal maximum likelihood estimate is

$$\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}). \tag{9}$$

The corresponding prediction error covariance matrix, $E[\tilde{x}_{k|k-1}\tilde{x}'_{k|k-1}]$, where $\tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1}$, is then given approximately by Eq. (8). The approximation again amounts to neglecting higher-order moments.

For $k=0$, we have

$$p(x_0|Z_{-1}) = p(x_0)$$

since Z_{-1} is vacuous. Consequently,

$$\hat{x}_{0|-1} = 0 \tag{10}$$

and

$$P_{0|-1} = P_0. \tag{11}$$

Equations (10) and (11) provide the initial conditions for prediction and filtering as we shall see in the sequel.

We remark that Eq. (9) is a necessary condition for marginal maximum likelihood prediction. It is also clearly a sufficient condition if $P_{k|k-1}$ is positive definite.

Turning now to the filtering problem, we have

$$\begin{aligned} L(x_k,Z_k) &= \log p(x_k|Z_k) \\ &= \log p(x_k,Z_k) - \log p(Z_k). \end{aligned}$$

But,

$$\begin{aligned} p(x_k,Z_k) &= p(x_k,z_k,Z_{k-1}) \\ &= p(z_k|x_k,Z_{k-1})p(x_k,Z_{k-1}) \\ &= p(z_k|x_k)p(x_k|Z_{k-1})p(Z_{k-1}) \end{aligned}$$

where use has been made of the fact that the conditional random vector z_k given x_k is independent of Z_{k-1} since $\{v_i,\ i=0,1,\ldots\}$ is an independent sequence.

Obviously, the only part of $L(x_k,Z_k)$ which depends on x_k is the product $p(z_k|x_k)p(x_k|Z_{k-1})$.

Since $\{v_i,\ i=0,1,\ldots\}$ is an independent, zero mean, gaussian, random sequence

$$p(z_k|x_k) = \eta_k \exp[-\tfrac{1}{2}||z_k - h(x_k)||^2_{R_k^{-1}}] \tag{12}$$

where

$$\eta_k = 1/\sqrt{(2\pi)^m |R_k|}\ .$$

It now follows from Eqs. (7), (9), and (12), and the above remark concerning $L(x_k,Z_k)$, that the marginal maximum likelihood filtered estimate is that value of x_k which minimizes

$$J(x_k,Z_k) = || z_k-h(x_k) ||^2_{R_k^{-1}} + || x_k-\hat{x}_{k|k-1} ||^2_{P_{k|k-1}^{-1}} .$$

Setting the gradient of $J(x_k,Z_k)$ with respect to x_k equal to zero yields the necessary condition

$$P_{k|k-1}^{-1}(x_k-\hat{x}_{k|k-1}) - h_x'(x_k)R_k^{-1}[z_k-h(x_k)] = 0 \tag{13}$$

where $h_x(x_k)$ is the $m \times m$ Jacobian matrix of $h(x_k)$. Equivalently, we have

$$x_k - \hat{x}_{k|k-1} - P_{k|k-1}h_x'(x_k)R_k^{-1}[z_k-h(x_k)] = 0. \tag{14}$$

Defining

$$g(x_k,z_k) = h_x'(x_k)R_k^{-1}[z_k-h(x_k)] \tag{15}$$

we have that Eqs. (13) and (14) are also sufficient conditions for the marginal maximum likelihood filtered estimate if

$$P_{k|k-1}^{-1} - g_x(x_k,z_k)$$

or equivalently

$$I - P_{k|k-1}g_x(x_k,z_k)$$

is positive definite where $g_x(x_k,z_k)$ is the $n \times n$ Jacobian matrix of $g(x_k,z_k)$ evaluated at that value of x_k which satisfies either Eq. (13) or Eq. (14), and I is the $n \times n$ identity matrix.

In order to obtain a sequential algorithm for filtering, we expand Eq. (15) about $x_k = \hat{x}_{k|k-1}$ to obtain

$$g(x_k,z_k) = g(\hat{x}_{k|k-1},z_k) + g_x(\hat{x}_{k|k-1},z_k)(x_k-\hat{x}_{k|k-1}).$$

Substituting this result into Eq. (14) and grouping terms, we get

$$[I-P_{k|k-1}g_x(\hat{x}_{k|k-1},z_k)]x_k - [I-P_{k|k-1}g_x(\hat{x}_{k|k-1},z_k)]\hat{x}_{k|k-1} - P_{k|k-1}g(\hat{x}_{k|k-1},z_k) = 0.$$

It now follows that an approximate solution of Eq. (14) for the desired filtered estimate of x_k is

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + [I-P_{k|k-1}g_x(\hat{x}_{k|k-1})]^{-1}P_{k|k-1}h_x'(\hat{x}_{k|k-1})R_k^{-1}[z_k-h(\hat{x}_{k|k-1})] \tag{16}$$

for $k=0,1,\ldots$. Equation (16) is evidently subject to the initial condition in Eq. (10).

After some computation, it can be shown that the covariance matrix $P_{k|k}$ of the filtering error $\tilde{x}_{k|k} = x_k - \hat{x}_{k|k}$ is given *approximately* by the expression

$$P_{k|k} = [I-P_{k|k-1}g_x(\hat{x}_{k|k-1},z_k)]^{-1}P_{k|k-1} \tag{17}$$

whose initial condition is specified by Eq. (11).

The desired sequential prediction-filtering algorithm is now completely defined by Eqs. (8)-(11), (16), and (17). This result is identical to that of Aoki [1] who expanded $p(x_k|Z_k)$ to obtain an approximation for the conditional mean $E[x_k|Z_k]$. Aoki treated only the scalar result, but his method can readily be utilized in the vector case. The present result also agrees with the invariant imbedding algorithm of Sage and Masters [2] which was obtained from $p(X_k|Z_k)$ where $X_k = \{x_0,x_1,\ldots,x_k\}$. Since the approximations utilized in the three developments are the same, it is no surprise that the resulting algorithms are also.

If it is assumed that there is no system disturbance in Eq. (1), the above algorithm is also identical to that of Pearson [3] who adopted the least-squares viewpoint. In this case, R_k^{-1} is an arbitrary weighting matrix and the filtered estimate of x_k is determined so that

$$J_k = \sum_{i=0}^{k} \| z_i-h(x_i) \|^2_{R_i^{-1}}$$

is minimized.

Finally, before proceeding to the smoothing problem, we point out an interesting and potentially useful possibility in connection with Eq. (14). Although this possibility is inherent in the results of Aoki [1], Sage and Masters [2], and Pearson [3], it is not explicit in their developments, and, we believe, has not been pointed out previously. Namely, we note that Eq. (14) is a system of n nonlinear algebraic equations to which Newton's method can be applied to obtain an iterative algorithm for filtering. In fact, Eq. (16) is the first step in the Newton process wherein the initial iterate is $x_k^1 = \hat{x}_{k|k-1}$. The general relation for the successive iterates is readily shown to be

$$x_k^{i+1} = [I-P_{k|k-1}g_x(x_k^i,z_k)]^{-1}\{\hat{x}_{k|k-1} + P_{k|k-1}[g(x_k^i,z_k)-g_x(x_k^i,z_k)x_k^i]\}$$

where $i=1,2,\ldots$ is the iteration index and all of the other terms have been defined previously.

We shall not pursue this method here, but we do note the need to develop a relation for the covariance matrix of the filtering error associated with each iterate. In addition, although sequential, the algorithm will, in general, not be useful for real-time filtering. Its potential in practice will, of course, depend upon the extent to which the iterative filtered estimates are "better" than those obtained via the algorithm of Eqs. (8)-(11), (16), and (17).

We consider the likelihood function

$$L(x_k,x_{k+1},Z_N) = \log p(x_k,x_{k+1}|Z_N)$$
$$= \log p(x_k,x_{k+1},Z_N) - \log p(Z_N)$$

for all $0 \le k < N$.

We observe that

$$p(x_k,x_{k+1},Z_N) = p(x_k,x_{k+1},z_{k+1},\cdots,z_N|Z_k)p(Z_k)$$

and that

$$\begin{aligned} p(x_k,x_{k+1},z_{k+1},\cdots,z_N|Z_k) &= p(x_{k+1},z_{k+1},\cdots,z_N|x_k,Z_k)p(x_k|Z_k) \\ &= p(x_{k+1},z_{k+1},\cdots,z_N|x_k)p(x_k|Z_k) \\ &= p(z_{k+1},\cdots,z_N|x_{k+1},x_k)p(x_{k+1}|x_k)p(x_k|Z_k) \\ &= p(z_{k+1},\cdots,z_N|x_{k+1})p(x_{k+1}|x_k)p(x_k|Z_k). \end{aligned} \tag{18}$$

Clearly, the only part of $L(x_k,x_{k+1},Z_N)$ which depends on x_k and x_{k+1} is the product in the last line of Eq. (18). The last two factors in this product are given by Eqs. (4) and (5), respectively, with k replaced by $k+1$.

We can therefore write

$$\max_{x_k,x_{k+1}} L(x_k,x_{k+1},Z_N)$$
$$= \max_{x_k,x_{k+1}} \{\log p(z_{k+1},\cdots,z_N|x_{k+1}) - \tfrac{1}{2}||x_k-\hat{x}_{k|k}||^2_{P^{-1}_{k|k}}$$
$$- \tfrac{1}{2}||x_{k+1}-f(x_k)||^2_{Q_k^{-1}}\}.$$

Assuming for the moment that $\hat{x}_{k+1|N}$ is given (which it is "initially" for $k=N-1$), we see that $\hat{x}_{k|N}$ is that value of x_k which minimizes

$$J(x_k,Z_N) = ||x_k-\hat{x}_{k|k}||^2_{P^{-1}_{k|k}} + ||\hat{x}_{k+1|N}-f(x_k)||^2_{Q_k^{-1}}.$$

The term $\log p(z_{k+1},\cdots,z_N|x_{k+1})$ does not appear in $J(x_k,Z_N)$ since it does not depend on x_k.

Equating the gradient of $J(x_k,Z_N)$ with respect to x_k to zero, we get the necessary condition

$$P^{-1}_{k|k}(x_k-\hat{x}_{k|k})-f'_x(x_k)Q_k^{-1}[\hat{x}_{k+1|N}-f(x_k)] = 0$$

which we rewrite as

$$x_k - \hat{x}_{k|k} - P_{k|k} f'_x(x_k) Q_k^{-1} [\hat{x}_{k+1|N} - f(x_k)] = 0. \tag{19}$$

Equation (19) has the same form as Eq. (14) for filtering. We therefore proceed to obtain an approximate algorithm for sequential smoothing in the same manner as we did for filtering in Section 3.

We set

$$\gamma(x_k, \hat{x}_{k+1|N}) = f'_x(x_k) Q_k^{-1} [\hat{x}_{k+1|N} - f(x_k)] \tag{20}$$

and note that Eq. (19) is a sufficient condition for determining $\hat{x}_{k|N}$ if either

$$P_{k|k}^{-1} - \gamma_x(x_k, \hat{x}_{k+1|N})$$

or

$$I - P_{k|k} \gamma_x(x_k, \hat{x}_{k+1|N})$$

is positive definite. Here, $\gamma_x(x_k, \hat{x}_{k+1|N})$ is the $n \times n$ Jacobian matrix of $\gamma(x_k, \hat{x}_{k+1|N})$ evaluated for x_k as determined from Eq. (19).

Expanding $\gamma(x_k, \hat{x}_{k+1|N})$ about $x_k = \hat{x}_{k|k}$ and retaining only first-order terms, we have

$$\gamma(x_k, \hat{x}_{k+1|N}) = \gamma(\hat{x}_{k|k}, \hat{x}_{k+1|N}) + \gamma_x(\hat{x}_{k|k}, \hat{x}_{k+1|N})(x_k - \hat{x}_{k|k}).$$

Substituting this result into Eq. (19), grouping terms, and solving for x_k, we obtain the following sequential relation for $\hat{x}_{k|N}$:

$$\hat{x}_{k|N} = \hat{x}_{k|k} + [I - P_{k|k} \gamma_x(\hat{x}_{k|k}, \hat{x}_{k+1|N})]^{-1} P_{k|k} f'_x(\hat{x}_{k|k}) Q_k^{-1} [\hat{x}_{k+1|N} - \hat{x}_{k+1|k}] \tag{21}$$

where we have substituted for $\gamma(\hat{x}_{k|k}, \hat{x}_{k+1|N})$ and utilized the fact that $f(\hat{x}_{k|k}) = \hat{x}_{k+1|k}$.

It is obvious that Eq. (21) defines a reverse-time sequential algorithm for which the starting conditions are $\hat{x}_{N|N}$, $\hat{x}_{N|N-1}$, and $\hat{x}_{N-1|N-1}$ when $k=N-1$. Implementation of Eq. (21) clearly requires that $\hat{x}_{k+1|k}$ and $\hat{x}_{k|k}$ be given for all $k=N-1,N-2,\ldots,0$.

After some computation, it can be shown that the covariance matrix $P_{k|N}$ of the smoothing error $\tilde{x}_{k|N} = x_k - \hat{x}_{k|N}$ is given *approximately* by the reverse-time sequential matrix equation

$$P_{k|N} = P_{k|k} - A_k (P_{k+1|k} - P_{k+1|N}) A'_k \tag{22}$$

where

$$A_k = [I - P_{k|k} \gamma_x(\hat{x}_{k|k}, \hat{x}_{k+1|N})]^{-1} P_{k|k} f'_x(\hat{x}_{k|k}) Q_k^{-1} \tag{23}$$

for $k=N-1,N-2,\ldots,0$. The required starting conditions for Eq. (22) are obviously $P_{N|N}$, $P_{N|N-1}$, and $P_{N-1|N-1}$, and the dependence of Eqs. (22) and (23) upon the prediction-filtering solution for input data in order to compute $P_{k|N}$ sequentially is equally clear.

Analogous to the situation in filtering, Eq. (19) is a result to which Newton's method can be applied to yield an iterative algorithm for smoothing. Also, as before, Eq. (21) defines the first step in the Newton process when $\hat{x}_{k|k}$ is chosen as the first iterate.

EXAMPLE

We consider the second-order system

$$(x_1)_{i+1} = \frac{(x_1)_i}{1+(x_1)_i(x_2)_i} + (w_1)_i$$

$$(x_2)_{i+1} = (x_2)_i + (w_2)_i$$

with the scalar measurement

$$z_j = (x_1)_j^3 + v_j$$

where $i=0,\ldots,9$ and $j=0,\ldots,10$. We take

$$P_0 = \begin{bmatrix} 25 & 0 \\ 0 & 25\cdot 10^{-4} \end{bmatrix} \qquad Q_i = \begin{bmatrix} 25\cdot 10^{-4} & 0 \\ 0 & 10^{-6} \end{bmatrix}$$

and $R_j = 10^{-2}$.

The true values of $(x_1)_{i+1}$, $(x_2)_{i+1}$, and z_j were generated taking $(x_1)_0 = 20$ and $(x_2)_0 = 0.3$, and utilizing the appropriate random number generators.

Filtering and smoothing were performed using the algorithms of Sections 3 and 4, respectively, with the initial estimates taken as $(\hat{x}_k)_{0|-1} = 18$ and $(\hat{x}_2)_{0|-1} = 0.34$. The resulting estimation error time histories are given below in Table 1 and the corresponding *approximate* error variances are shown in Table 2.

i	$\tilde{x}_1$		$\tilde{x}_2$	
	Filter	Smooth	Filter	Smooth
0	$-.9655$	$-.9655$	$-.4000\cdot 10^{-1}$	$.4533\cdot 10^{-2}$
2	$-.4060\cdot 10^{-2}$	$.5738\cdot 10^{-2}$	$.9011\cdot 10^{-2}$	$.7135\cdot 10^{-2}$
4	$-.3824\cdot 10^{-1}$	$-.3891\cdot 10^{-1}$	$.9325\cdot 10^{-2}$	$.9141\cdot 10^{-2}$
6	$-.6643\cdot 10^{-1}$	$-.5388\cdot 10^{-1}$	$.9868\cdot 10^{-2}$	$.9686\cdot 10^{-2}$
8	$-.6807\cdot 10^{-1}$	$-.7184\cdot 10^{-1}$	$.1032\cdot 10^{-1}$	$.1036\cdot 10^{-1}$
10	$-.2203\cdot 10^{-1}$	$-.2203\cdot 10^{-1}$	$.1074\cdot 10^{-1}$	$.1074\cdot 10^{-1}$

Table 1

Filtering and smoothing error time histories.

i	P_{11}		P_{22}	
	Filter	Smooth	Filter	Smooth
0	$.1407 \cdot 10^{-7}$	$.1407 \cdot 10^{-7}$	$.2500 \cdot 10^{-2}$	$.2739 \cdot 10^{-4}$
2	$.1775 \cdot 10^{-3}$	$.1749 \cdot 10^{-3}$	$.5144 \cdot 10^{-4}$	$.4972 \cdot 10^{-4}$
4	$.1193 \cdot 10^{-2}$	$.1125 \cdot 10^{-2}$	$.5190 \cdot 10^{-4}$	$.5170 \cdot 10^{-4}$
6	$.2678 \cdot 10^{-2}$	$.2416 \cdot 10^{-2}$	$.5375 \cdot 10^{-4}$	$.5369 \cdot 10^{-4}$
8	$.3822 \cdot 10^{-2}$	$.3752 \cdot 10^{-2}$	$.5570 \cdot 10^{-4}$	$.5569 \cdot 10^{-4}$
10	$.5847 \cdot 10^{-2}$	$.5847 \cdot 10^{-2}$	$.5769 \cdot 10^{-4}$	$.5769 \cdot 10^{-4}$

Table 2

Approximate filtering and smoothing error variances.

The data here, of course, represent only one sample function of the algorithms' behavior and should be interpreted accordingly. A statistical analysis based on an ensemble of runs is necessary to obtain a realistic picture of performance. Thus, the data in the two tables are, at best, only indicative of actual performance. Further, the estimates are, in general, biased. Computation of this bias is of obvious importance.

REFERENCES

1. M. Aoki, "Optimization of Stochastic Systems," Academic Press, Inc., New York, 1967.

2. A. P. Sage and G. W. Masters, "On-Line Estimation of States and Parameters for Discrete Non-Linear Dynamic Systems," *Proceedings of the National Electronics Conference*, vol. 22, pp. 677-682, 1966.

3. J. B. Pearson, "On Nonlinear Least-Squares Filtering," *Automatica*, 4 (1967), 97-105.

4. W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. II, John Wiley and Sons, New York, 1966.

A SECOND ORDER FILTER FOR STATE ESTIMATION IN NONLINEAR DISCRETE SYSTEMS

CHAW-BING CHANG and HINRICH R. MARTENS
Department of Electrical Engineering
State University of New York at Buffalo
Buffalo, New York 14214

ABSTRACT

A second order filter for the estimation of state variables in nonlinear discrete systems has been derived. The approach used is based upon Bayes' Rule.

Conventional models for the nonlinear system and measurement equations are adopted while Gaussian uncorrelated random sequences are assumed for the system noise and measurement noise.

Second order terms are introduced through a Taylor series expansion of the nonlinear dynamic and measurement equations.

The use of Bayes' Rule requires the evaluation of conditional probability density functions and covariances involving state variables and measurements in terms of the estimate of the state variables. Several lemmas are utilized to obtain simplified expressions.

The resulting filter equations are conformant in structure to the linear Kalman filter with the second order terms appearing as additive corrective terms. The filter is therefore readily applied.

Numerical results of test problem indicate greatly improved performance over filters including only first order terms.

I. INTRODUCTION

The purpose of this paper is to derive a discrete second order nonlinear estimation filter. The derivation follows the Bayesian approach as first proposed by Lee [1]. Ho and Lee [2] have also utilized the Bayesian approach for the solution of the general estimation problems and provided explicit recursive relations for the linear problem which are identical to the Kalman filter equations [3]. The estimation of states of nonlinear systems is made possible by an extension of linear theory through the use of approximations. First-order approximations result in various forms of a so-called quasilinear Kalman filter or first-order filter as developed by Cox [4], Friedland and Bernstein [5], Detchmendy and Sridhar [6], Richmond and Thau [7]. Simulation results generally indicate that a quasilinear filter will provide state estimation of acceptable quality for systems with minor nonlinearities. But when estimation of state variables in systems with severe nonlinearities (square roots, transcendental relations) is desired either in the system dynamics or the measurement equations it is frequently the experience that first-order algorithms lead to diverging, or at best, non-acceptable results. In these cases it has been demonstrated that a second-order filter is greatly more effective. The development of second-order filters for state estimation has been carried out by Athans, Wishner and Bertolini [8]. Second order Taylor series expansions of the system dynamics and the measurement equations are included in the algorithm defining the filter. Sage and Master [9] have derived a discrete filter involving higher order terms through an invariant imbedding approach, Simulation results in both cases support the fact that a second order filter is superior in state estimation of severely nonlinear system.

In this paper, the problem formulation and the Bayesian approach to the filter problem are first outlined in section II. The filter equations are subsequently derived in section III. An example dealing with the estimation of the trajectory of a free falling body which follows a ballistic trajectory is then solved by both a first order and the proposed second order filter. The continuous dynamics of this free falling body is sampled to form a discrete equivalence. The derived second order filter can then be readily applied.

II. PROBLEM FORMULATION AND PROPOSED APPROACH

The problem to be considered here is the estimation of x_k of the nonlinear system

$$x_{k+1} = f(x_k) + Gu_k \tag{1}$$

from the observation equation

$$y_{k+1} = h(x_{k+1}) + v_{k+1} \tag{2}$$

where x_k, u_k, y_k and v_k are n, m, p, and p vectors respectively. The variables u_k and v_k are white random disturbances with Gaussian distribution and zero mean. The optimum estimate $\hat{x}_k$ of x_k is defined as the conditional expection of x_k given the observations Y_k, where Y_k is the vector space spanned by $y_0, y_1, y_2, \ldots, y_k$, i.e.

$$\hat{x}_k = E[x_k/Y_k] \tag{3}$$

On the basis of the assumptions made, one may write

$$p(x_k/Y_k) = \frac{1}{(2\pi)^{n/2}|P_k|^{1/2}} \exp\left\{-1/2[x_k-\hat{x}_k]^T P_k^{-1}[x_k-\hat{x}_k]\right\} \tag{4}$$

where P_k is the conditional covariance defined by

$$P_k = \mathrm{Cov}[x_k/Y_k] = E\left\{(x_k-\hat{x}_k)(x_k-\hat{x}_k)^T/Y_k\right\} \tag{5}$$

If one substitutes (4) and similarly corresponding expressions into the equation defining Bayes' rules, i.e.

$$p(x_{k+1}/Y_{k+1}) = \frac{p(y_{k+1}/x_{k+1})}{p(y_{k+1}/Y_k)}\ p(x_{k+1}/Y_k) \tag{6}$$

it is possible to express the right hand side of (6) entirely in terms of $\hat{x}_k$. A set of recursive equations may then be obtained for estimating the state variable x_k.

The approach just outlined is the so called Bayesian Approach to estimation problems. An exact and readily applicable recursive formula is only available for linear systems. In order to obtain a solution for nonlinear systems an exact solution may be approximated by linearizing the dynamics and measurements by a Taylor series expansion about the extrapolated states and neglecting the second and higher order terms. It is obvious that the resulting linear equations can be fairly inaccurate in the sense of approximating a nonlinear system when the system is severely nonlinear. The task of retaining more terms in the Taylor series expansion is then necessary for the purpose of improving the suboptimal nonlinear filter performance. The filter equations to be derived in this paper will retain the second order term.

The Taylor series expansions of (1) and (2) are stated as follows:

$$x_{k+1} \approx f(\hat{x}_k) + F(x_k-\hat{x}_k) + \frac{1}{2}\sum_{i=1}^{n} \varphi_i (x_k-\hat{x}_k)^T I_i (x_k-\hat{x}_k) + Gu_k \tag{7}$$

$$y_{k+1} \approx h(\tilde{x}_{k+1}) + H(x_{k+1}-\tilde{x}_{k+1}) + \frac{1}{2}\sum_{i=1}^{p} \varphi_i (x_{k+1}-\tilde{x}_{k+1})^T$$

$$\cdot J_i(x_{k+1} - \tilde{x}_{k+1}) + V_{k+1} \tag{8}$$

where $\tilde{x}_{k+1}$ is an extrapolated state defined by

$$\tilde{x} \triangleq E[x_{k+1}/Y_k] \tag{9}$$

and the matrices are

$$F = [\nabla_x f^T(x_k)]^T \Big|_{x_k = \hat{x}_k}$$

$$H = [\nabla_x h^T(x_{k+1})]^T \Big|_{x_{k+1} = \tilde{x}_{k+1}}$$

$$\left.\begin{aligned} [I_i]_{L,M} &= \frac{\partial^2 f_i}{\partial x_L \partial x_M}\Big| \quad x_k = \hat{x}_k \\ [J_i]_{L,M} &= \frac{\partial^2 h_i}{\partial X_L \partial x_M}\Big| \quad x_{k+1} = \tilde{x}_{k+1} \end{aligned}\right\} \quad \begin{aligned} L &= 1, 2, \ldots n \\ M &= 1, 2, \ldots n \end{aligned}$$

$$\varphi_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{with "1" in the } i^{th} \text{ component.}$$

The following lemma is used in the ensuing developments.

Lemma [8]. Given

(1) a Gaussian random variable z with zero mean

(2) $P = E[zz^T/Y]$

(3) I_1 and I_2 are symmetric positive definite matrices Then

$\Rightarrow$(1) $E[z^T I_1 Z/Y] = Tr[I_1 P]$

(2) $E[zTr[I_1 zz^T]] = 0$

(3) $E[z^T I_1 zz^T I_2 Z/Y] = 2Tr[I_1 P I_2 P] + Tr[I, P]Tr[I_2 P]$

are true.
where Tr[A] indicates the trace of matrix A.

III. DERIVATION OF FILTER EQUATIONS

Three conditional density functions need to be evaluated in (6), i.e.,

$p(y_{k+1}/x_{k+1})$, $p(x_{k+1}/Y_k)$ and $p(y_{k+1}/Y_k)$.

Because of the Gaussian distribution Equation (4) may be used in their evaluation. It becomes therefore necessary to determine the conditional mean and covariance of each of the arguments involved.

Let the covariance matrices of u_k and v_k be denoted by Q_k and R_k respectively. The required terms are then evaluated as follows.

(1) for $p(x_{k+1}/Y_k)$:

$$E[x_{k+1}/Y_k]$$
$$= E\left\{[\, f(\hat{x}) + F(x-\hat{x}) + \frac{1}{2}\sum_{i=1}^{n} \varphi_i (x-\hat{x})^T I_i (x-\hat{x}) + Gu]/Y\right\}$$
$$= f(\hat{x}) + \frac{1}{2}\sum_{i=1}^{n} \varphi_i \mathrm{Tr}[I_i P]$$
$$\triangleq \tilde{x}_{k+1} \qquad (10)$$

$$\mathrm{Cov}[x_{k+1}/Y]$$
$$= E\left\{[F\cdot(x_k-\hat{x}_k) + \frac{1}{2}\sum_{i=1}^{n} \varphi_i (x_k-\hat{x}_k)^T I_i (x_u-\hat{x}_k) + Gu_k - \frac{1}{2}\sum_{i=1}^{n} \varphi_i \mathrm{Tr}[I_i P]]\right.$$
$$\left.\cdot\,[F\cdot(x_k-\hat{x}_k) + \frac{1}{2}\sum_{i=1}^{n} \varphi_i (x_k-\hat{x}_k)^T I_i (x_k-\hat{x}_k) + Gu_k - \frac{1}{2}\sum_{i=1}^{n} \varphi_i \mathrm{Tr}(I_i P)]^T /Y_k\right\}$$
$$= FP_kF^T + GQ_kG^T + \frac{1}{2}S_k$$
$$\triangleq M_{k+1} \qquad (11)$$

where

$$S = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi_i \mathrm{Tr}[I_i P I_j P]\, \varphi_j^T$$

(2) For $p(y_{k+1}/x_{k+1})$:

$$E[y_{k+1}/x_{k+1}]$$
$$= E[(h(x_{k+1}) + u_{k+1})/x_{k+1}]$$
$$= h(x_{k+1}) \qquad (12)$$

$$\mathrm{Cov}[y_{k+1}/x_{k+1}]$$

$$= E\left\{[y_{k+1}-h(x_{k+1})][y_{k+1}-h(x_{k+1})]^T/x_{k+1}\right\}$$

$$= R_{k+1} \tag{13}$$

(3) For $p(y_{k+1}/Y_k)$:

$$E(y_{k+1}/Y_k)$$

$$= E\left\{\left[h(\tilde{x}_{k+1}) + H[x_{k+1}-\tilde{x}_{k+1}] + \frac{1}{2}\sum_{i=1}^{p}\varphi_i(x_{k+1}-\tilde{x}_{k+1})^T J_i(x_{k+1}-\tilde{x}_{k+1})\right]/Y\right\}$$

$$= h(\tilde{x}_{k+1}) + \frac{1}{2}\sum_{i=1}^{n}\varphi_i \mathrm{Tr}[J_i M_{k+1}] \tag{14}$$

$$\mathrm{Cov}[y_{k+1}/Y_k]$$

$$= E\left\{[H\cdot(x_{k+1}-\tilde{x}_{k+1}) + \frac{1}{2}\sum_{i=1}^{p}\varphi_i(x_{k+1}-\tilde{x}_{k+1})^T J_i(x_{k+1}-\tilde{x}_{k+1}) + V_{k+1} -\right.$$

$$\left.\frac{1}{2}\sum_{i=1}^{p}\varphi_i \mathrm{Tr}[J_i M_{k+1}]\right]\cdot\left[H(x_{k+1}-\tilde{x}_{k+1}) + \frac{1}{2}\sum_{i=1}^{p}\varphi_i(x_{k+1}-\tilde{x}_{k+1})^T J_i(x_{k+1}-\tilde{x}_{k+1}) + V_{k+1}\right.$$

$$\left.-\frac{1}{2}\sum_{i=1}^{p}\varphi_i \mathrm{Tr}(J_i M_{k+1})\right]^T/Y_k\Big\}$$

$$= HM_{k+1}H^T + R_{k+1} + \frac{1}{2}W_{k+1} \tag{15}$$

where

$$W_{k+1} = \sum_{i=1}^{p}\sum_{j=1}^{p}\varphi_i \mathrm{Tr}[J_i M_{k+1} J_j M_{k+1}]\varphi_j^T$$

Substituting these results into the right hand side of equation (6) yields

$$\frac{p(y_{k+1}/x_{k+1})}{p(y_{k+1}/Y_k)}\; p(x_{k+1}/Y_k)$$

$$= \frac{|HM_{k+1}H^T+R_{k+1}+\frac{1}{2}W_{k+1}|^{1/2}}{(2\pi)^{n/2}|R_{k+1}|^{1/2}|M_{k+1}|^{1/2}} \cdot \exp\Big\{ -1/2[(x_{k+1}-\tilde{x}_{k+1})^T \cdot M_{k+1}^{-1} \cdot$$

$$(x_{k+1}-\tilde{x}_{k+1}) - \frac{1}{2}[(y_{k+1}-h(x_{k+1}))^T \cdot R_{k+1}^{-1} \cdot (y_{k+1}-h(x_{k+1}))] +$$

$$\frac{1}{2}[(y_{k+1}-h(\tilde{x}_{k+1}) - \frac{1}{2}\sum_{i=1}^{p}\varphi_i Tr[JiM_{k+1}])^T \cdot [HMH^T+R+\frac{1}{2}W]^{-1} \cdot$$

$$(y_{k+1}-h(\tilde{x}_{k+1}) - \frac{1}{2}\sum_{i=1}^{p}\varphi_i[J_iM_{k+1}])]\Big\} \qquad (16)$$

Multiplying out the arguments of the exponential terms and comparing with the exponent of the left hand side exponents of equation (6) which are

$$-\frac{1}{2}[(x_{k+1}-\hat{x}_{k+1})^T P_{k+1}^{-1}(x_{k+1}-\hat{x}_{k+1})]$$

$$= -\frac{1}{2}\Big\{x_{k+1}^T P_{k+1}^{-1}x_{k+1} - x_{k+1}^T P_{k+1}^{-1}\hat{x}_{k+1} - \hat{x}_{k+1}^T P_{k+1}^{-1}x_{k+1} + \hat{x}_{k+1}^T P_{k+1}^{-1}\hat{x}_{k+1}\Big\} \qquad (17)$$

one has the following results

(1) Quadratic terms in X_{k+1}

$$x_{k+1}^T P_{k+1}^{-1}x_{k+1}$$

$$= x_{k+1}^T M^{-1}x_{k+1} + x_{k+1}^T H^T R_{k+1}^{-1}Hx_{k+1} + x_{k+1}^T [\sum_{i=1}^{p}$$

$$(-y_{k+1}^T + h^T(\tilde{x}_{k+1}))R^{-1}\varphi_i J_i] \cdot X_{k+1}$$

This implies that

$$P_{k+1}^{-1} = M_{k+1}^{-1} + H^T R_{k+1}^{-1}H + \sum_{i=1}^{p}(-y_{k+1}^T + h^T(\tilde{x}_{k+1}))R^{-1}\varphi_i J_i$$

By using the Matrix Inversion Lemma this yeilds

$$P_{k+1} = M^*_{k+1} - M^*_{k+1} H^T [H M^*_{k+1} H^T + R_{k+1}]^{-1} H M^*_{k+1} \tag{18}$$

where

$$M^*_{k+1} = [M^{-1}_{k+1} + \sum_{i=1}^{p} (-y^T_{k+1} + h^T(\tilde{x}_{k+1})) R^{-1}_{k+1} \varphi_i J_i]^{-1} \tag{19}$$

(2) First order terms in $\bar{x}_{k+1}$

$$x^T_{k+1} P^{-1}_{k+1} \hat{x}_{k+1}$$

$$= X^T_{k+1} M^{-1}_{k+1} \tilde{x}'_{k+1} + x^T_{k+1} H^T R_{k+1} y_{k+1} - x^T_{k+1} H^T R_{k+1} h(\tilde{x}_{k+1})$$

$$- x^T_{k+1} (\sum_{i=1}^{p} y^T_{k+1} R^T_{k+1} \varphi_i J_i) \tilde{x}_{k+1} + x^T_{k+1} (\sum_{i=1}^{p} h^T(\tilde{x}_{k+1}) R^T_{k+1} \varphi_i J_i) \tilde{x}_{k+1}$$

$$+ x^T_{k+1} H^T R^{-1}_{k+1} H \tilde{x}_{k+1}$$

from which it follows that

$$P^{-1}_{k+1} \hat{x}_{k+1} = M^{-1}_{k+1} \tilde{x}_{k+1} + H^T R^{-1}_{k+1} y_{k+1} - H^T R^{-1}_{k+1} h(\tilde{x}_{k+1})$$

$$- (\sum_{i=1}^{p} y^T_{k+1} R^{-1} \varphi_i J_i) \tilde{x}_{k+1} + (\sum_{i=1}^{p} h^T(\tilde{x}_{k+1}) R^T_{k+1} \varphi_i J_i) \tilde{x}_{k+1} + H^T R^{-1}_{k+1} H \tilde{x}_{k+1}$$

Premultiplying the above equation by P_{k+1} and using equations (18) and (19) one has

$$\hat{x}_{k+1} = \tilde{x}_{k+1} + P_{k+1} H^T R^{-1} (y_{k+1} - h(\tilde{x}_{k+1})) \tag{20}$$

Combining equations (18), (19), and (20) with previously defined relations, one has the following summarized filter equations:

$$\hat{x}_{k+1} = \tilde{x}_{k+1} + P_{k+1} H^T_{k+1} R^{-1}_{k+1} [y_{k+1} - h(\tilde{x}_{k+1})] \tag{20}$$

$$\tilde{x}_{k+1} = f(\hat{x}_k) + \frac{1}{2} \sum_{i=1}^{n} \varphi_i T_r [I_i P] \tag{10}$$

$$P_{k+1} = M^*_{k+1} - M^*_{k+1} H^T_{k+1} [H_{k+1} M^*_{k+1} H^T_{k+1} + R_{k+1}]^{-1} H_{k+1} M^*_{k+1} \tag{18}$$

$$M^{*}_{k+1} = [M^{-1}_{k+1} + \sum_{i=1}^{p} (-y^{T}_{k+1} + h^{T}(\tilde{x}_{k+1}))R^{-1}_{k+1}\varphi_i J_i]^{-1} \tag{19}$$

$$M_{k+1} = F_k P_k F_k^{T} + G_k Q_k G_k^{T} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_i [Tr[I_i P_k I_j P_k]] \varphi_j^{T} \tag{11}$$

Note that if I_i and J_i are dropped, these equations are reduced to a first order filter. If the system is linear, they are identical to the equation of the Kalman filter.

IV. EXAMPLE †

To evaluate the performance of the derived filter, the filter equations are employed in a numerical example dealing with the estimation of state variables in a vertically falling body. The geometry of the problem is illustrated in Figure 1. Let the following variables be defined

x_1: altitude
x_2: velocity (downward)
m: mass (constant)
c_p: drag coefficient (constant)
A: projected area (constant)
p: air density
H: radar altitude
M: horizontal distance
r(t): range

If the effect of gravity is negligible, the equations of motion are

$$\left.\begin{aligned} \dot{x}_1(t) &= -x_2(t) \\ \dot{x}_2(t) &= -\frac{C_p A p}{2m} x_2^{2}(t) \end{aligned}\right\} \tag{21}$$

where

$$p = p_0 e^{-rx_1(t)} \quad , \quad r = 5 \times 10^{-5}$$

The ballistic parameter, defined as

$$x_3 \triangleq C_p A p_0 / 2m$$

usually cannot be known precisely. Under this assumption, the state equations are

$$\left.\begin{aligned} \dot{x}_1 &= -x_2 \triangleq f_1(x) \\ \dot{x}_2 &= -e^{-rx_1} x_2 x_3 \triangleq f_2(x) \\ \dot{x}_3 &= 0 \triangleq f_3(x) \end{aligned}\right\} \tag{22}$$

†This example solves the same problem of Athans, Wishner, and Bertolini [8].

The measurements are taken at discrete instants of time by a radar that measures range in the presence of white Gaussian noise, i.e.

$$r(t) = \sqrt{M^2 + [x_1 - H]^2} \tag{23}$$

The observed range y_k is the following equation

$$y_k = \sqrt{M^2 + [x_1(k) - H]^2} + V_k \tag{24}$$

In order to allow the continuous dynamics to be processed by the discrete filter, equation (22) is "discretized" by the following approximating formula

$$x_{k+1} = \int_{t_k}^{t_{k+1}} f(x)dt + x_k \tag{25}$$

The time interval $t_{k+1} - t_k$ corresponds to the radar data available.

The Runge-Kutte-Merson numerical integration technique is employed in this simulation. The following initial values and constants are chosen in this example.

(1) True initial values of state variables

$$\begin{aligned} x_1 &= 300000 \text{ ft.} \\ x_2 &= 20000 \text{ ft./s.} \\ x_3 &= 10^{-3} \end{aligned} \tag{26}$$

(2) Estimated initial value of state variables

$$\begin{aligned} x_1 &= 300000 \text{ ft.} \\ x_2 &= 20000 \text{ ft.} \\ x_3 &= 3 \times 10^{-5} \end{aligned} \tag{27}$$

(3) The initial error covariance matrix is selected as a diagonal matrix with the following elements

$$\begin{aligned} P_{11} &= 1000000 \\ P_{22} &= 4000000 \\ P_{33} &= 0.0001 \end{aligned} \tag{28}$$

(4) The radar altitude and horizontal distance are selected as

$$\begin{aligned} H &= 100,000 \text{ ft.} \\ M &= 100,000 \text{ ft.} \end{aligned} \tag{29}$$

(5) The radar observations are contaminated by the white Gaussian Noise with covariance $R = 10,000 \text{ ft}^2$

The aboslute value of estimation errors of x_1, x_2, and x_3 are plotted in Figures 2, 3, and 4 respectively. One can find that the retaining of second order terms in Taylor series expansion can markedly improve the filter performance.

(1) Lee, R. C. K., "Optimal Estimation, Identification and Control" Research Monagraph No. 26, The M. I. T. Press, Cambridge, Mass. 1964

(2) Ho, Y. C., abd Lee, R. C. K., "A Bayesian Approach to Problems in Stochastic Estimation and Control" IEEE Trans. Automatic Control, October 1964

(3) Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems" Trans ASME, Sec. D, Vol. 81D, 1960.
(4) Cox, H., "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems" IEEE Trans. Automatic Control Vol. AC-9, January 1964
(5) Friedland, B., and Berstein, I., "Estimation of the State of a Nonlinear Process in the Presence of Non-Gaussian Noise and Distributions" Journal of the Frankline Institute. Vol. 281, No. 6, June 1966
(6) Detchmendy, D. M.,, , and Sridhar, R. "Sequential Estimation of States and Parameters in Noisy Nonlinear Dynamical Systems" Trans. ASME June, 1966.
(7) Richman, J., and Than, F., "Elements of the Kalman Filtering Technique" TNB Research Bulletin, Vol. 8, 9, 1967
(8) Athans M., Wishner, R. P., and Bertolini, A. "Suboptimal State Estimation for Continuous-Time Nonlinear Systems from Discrete Noisy Measurements" IEEE Trans. Automatic Control Vol. AC-13, No. 5, October 1968

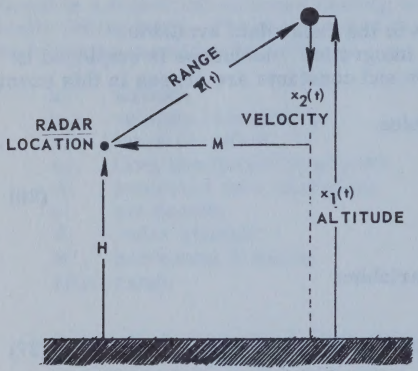


FIGURE 1 – The Geometry of the Example

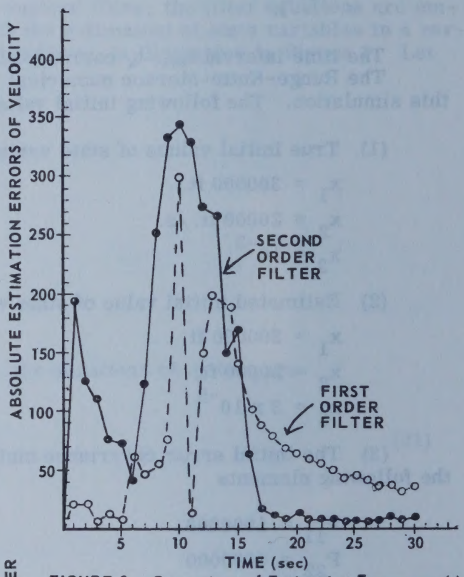


FIGURE 3 – Comparison of Estimation Errors – $x_2(t)$

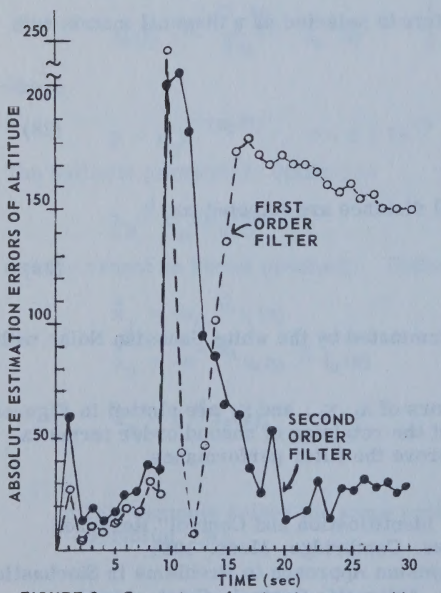


FIGURE 2 – Comparison of estimation errors $x_1(t)$

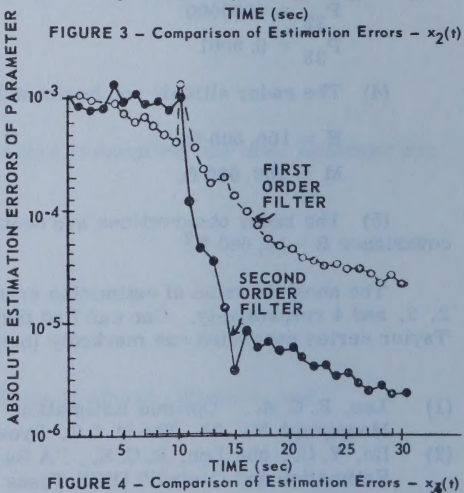


FIGURE 4 – Comparison of Estimation Errors – $x_3(t)$

STABILIZATION OF A CLASS OF NONLINEAR SYSTEMS USING A LINEAR SYSTEM ESTIMATOR*

ROBERT E. STRANE† WILLIAM G. VOGT
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania

ABSTRACT

The use of a linear asymptotic estimator to stabilize certain types of nonlinear systems is considered. The asymptotic estimator used is a linear dynamical system whose output state, $\underline{z}$, provides an asymptotic estimation of the state, $\underline{x}$, of the observed system. The output state, $\underline{z}$, is used to provide feedback which insures the absolute stability of the compensated system.

INTRODUCTION

In the theory of feedback control systems, the establishment of stability of the set of differential equations which describes the system, is the primary concern of the designer. In the case where the system, to be controlled is a completely linear stationary system, very powerful and explicit mathematical criteria have been established which allow both the analysis of an existing system and the synthesis of a stable control system from a given plant configuration. For nonlinear systems several techniques are available which give information concerning the stability of systems whose defining differential equations are known, however, the question of the synthesis of stable nonlinear systems by designing a compensator which insures the stability of the compensated nonlinear plant has remained relatively unexplored when compared to the synthesis of linear systems. In the following discussion it will be shown that for the class of nonlinear systems whose analysis was originally investigated by A. I. Lur'e [1] a compensator can be designed which insures the desired stability.

The Lur'e systems to be investigated are described by the differential equations

$$\begin{aligned} \dot{\underline{x}} &= \underline{A}\,\underline{x} + \underline{b}\,f(\sigma) + \underline{b}\,u \\ \sigma &= \underline{c}'\,\underline{x} \\ y_1 &= f(\sigma) \\ y_2 &= \underline{d}'\,\underline{x} \end{aligned} \tag{1}$$

where $\underline{x}$, $\underline{b}$, $\underline{c}$ and $\underline{d}$ are n-vectors, u is a single system input, y_1 and y_2 are available system outputs, and $f(\sigma)$ is a single valued, continuous real function, which is defined for all real values of its argument, σ, and which satisfies the conditions

$$f(0) = 0; \qquad 0 \leq \frac{f(\sigma)}{\sigma} \leq k \qquad \text{for } \sigma \neq 0$$

$$\lim_{\sigma \to \infty} \int_0^{\sigma} f(s)ds = \infty \tag{2}$$

---

* This research was supported in part by the National Aeronautics and Space Administration under Grant NSG 416 with the University of Pittsburgh and is a part of the Doctoral Dissertation of Robert E. Strane.

† Presently with Fecker Systems Division, Owens-Illinois, Inc., Pittsburgh, Pa.

For this system with $k < \infty$, and both outputs y_1 and y_2 available, it will be shown that if the pair $(\underline{A}, \underline{d}')$ is completely observable and the pair $[\underline{A},\underline{b}]$ is completely controllable, then a compensator can always be constructed for the system which will insure absolute stability of the compensated system. In the case where the only available output is the nonlinear element, y_1, then absolute stability can be achieved if $\underline{A}$ is a stable matrix. For cases where the pair $(\underline{A},\underline{d}')$ is not completely observable, then under certain conditions, but not all, the compensated nonlinear system can be insured to be absolutely stable.

In order to prove the above assertions it is necessary to make use of several well known theorems. These theorems are now presented for reference without proof.

Theorem 1: (Popov Stability Criterion [2]). For the system described by the set of differential equations given by equations (1) and (2) to be absolutely stable for nonlinearities contained in the sector [0,k] for systems where $\underline{A}$ has only eigenvalues containing non zero negative parts (principal cases) and in the sector $[\varepsilon,k]$ for systems with $\underline{A}$ having some purely imaginary eigenvalues (particular case) it is sufficient that their exist a finite real number, q, such that for all finite $w \geq 0$ the following inequality is satisfied

$$\text{Re } (1 + jqw)\ G(jw) + \frac{1}{k} > 0$$

where $G(s) = -\underline{c}'(\underline{Is}-\underline{A})^{-1}\underline{b}$. In addition, in particular cases, the condition for stability in the limit (i.e. where $f(\sigma) \to \varepsilon\sigma$, $\varepsilon > 0$) must be satisfied.

Theorem 2: (Lyapunov's Second Method [3]). Given the dynamic system

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{f}(\underline{x})$$

where

$$\underline{f}(\underline{0}) = \underline{0}.$$

Suppose there exists a scalar function $V(\underline{x})$ with continuous first partial derivatives with respect to $\underline{x}$ such that $V(\underline{0}) = 0$, and (a) $V(\underline{x})$ is positive definite; and (b) $\dot{V}(x)$, taken along the solutions of the system differential equations starting at any $\underline{x} \neq 0$ in some region Ω_o about the origin, is negative definite. Then the equilibrium state of the system, $\underline{x}=\underline{0}$, is asymptotically stable within Ω_o.

Theorem 3: [4]. Given a nxn constant matrix $\underline{A}$ and an nxl column matrix $\underline{b}$ where $[\underline{A},\underline{b}]$ represent a completely controllable pair. The eigenvalues of the matrix $\underline{A} + \underline{b}\,\underline{k}'$, where $\underline{k}'$ is a lxn row matrix, can be arbitrarily chosen by proper selection of the matrix $\underline{k}'$.

Theorem 4: [5]. Given a nxn constant matrix $\underline{A}$ and lxn row matrix $\underline{c}'$ where $[\underline{A},\underline{c}']$ represent a completely observable pair. The eigenvalues of the matrix $\underline{A} + \underline{k}\,\underline{c}'$, where $\underline{k}$ is an nxl column matrix, can be arbitrarily chosen by proper selection of the matrix $\underline{k}$.

STABILIZATION OF LUR'E TYPE SYSTEMS

Stabilization Using Both y_1 and y_2

Consider the system described by the set of differential equations (1) and (2). In addition assume that the pair $[\underline{A},\underline{d}']$ is a completely observable pair and the pair $[\underline{A}, \underline{b}]$ is a completely controllable pair. Now consider that an observer is constructed and placed in cascade with this system. The equations describing this observer are, as shown in Figure 1:

$$\dot{\underline{z}} = \underline{A}\,\underline{z} + \underline{b}\,y_1 + \underline{b}\,u + \underline{g}\,e$$
$$e = y_2 - \underline{d}'\,\underline{z} \tag{3}$$

This observer, sometimes called a linear system estimator, was first studied by Luenberger [5] and shown to provide an asymptotic estimation of the state, $\underline{x}$, of the observed system. If feedback from the output state, $\underline{z}$, of the observer is applied in the form of

$$u = \underline{k}'\,\underline{z}\ , \tag{4}$$

the resulting system differential equations are:

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,f(\sigma) + \underline{b}\,\underline{k}'\,\underline{z}$$
$$\dot{\underline{z}} = \underline{A}\,\underline{z} - \underline{g}\,\underline{d}'\,\underline{z} + \underline{b}\,\underline{k}'\,\underline{z} + \underline{g}\,\underline{d}'\,\underline{x} + \underline{b}\,f(\sigma) \tag{5}$$

where

$$\sigma = \underline{c}'\,\underline{x}$$

If (5) is expressed in matrix notation, the resulting system equations are

$$\begin{bmatrix}\dot{\underline{x}}\\ \dot{\underline{z}}\end{bmatrix} = \begin{bmatrix}\underline{A} & \underline{b}\,\underline{k}'\\ \underline{g}\,\underline{d}' & \underline{A}-\underline{g}\,\underline{d}'+\underline{b}\,\underline{k}'\end{bmatrix}\begin{bmatrix}\underline{x}\\ \underline{z}\end{bmatrix} + \begin{bmatrix}\underline{b}\\ \underline{b}\end{bmatrix} f(\sigma) \tag{6}$$

$$\sigma = [\underline{c}',\ \underline{0}]\begin{bmatrix}\underline{x}\\ \underline{z}\end{bmatrix}$$

When the similarity transformation

$$\underline{T}_1 = \begin{bmatrix}\underline{I} & \underline{0}\\ \underline{I} & -\underline{I}\end{bmatrix} \tag{7}$$

is applied to the compensated system matrix differential equations (6) the following representation of the system is obtained

$$\dot{\underline{w}} = \begin{bmatrix}\underline{A} + \underline{b}\,\underline{k}' & -\underline{b}\,\underline{k}'\\ \underline{0} & \underline{A}-\underline{g}\,\underline{d}'\end{bmatrix}\underline{w} + \begin{bmatrix}\underline{b}\\ \underline{0}\end{bmatrix} f(\sigma) \tag{8}$$

where

$$\underline{w} = \underline{T}_1\begin{bmatrix}\underline{x}\\ \underline{z}\end{bmatrix} = \begin{bmatrix}\underline{x}\\ \underline{x}-\underline{z}\end{bmatrix} \tag{9}$$

The output, σ, is given by

$$\sigma = [\underline{c}',\ \underline{0}]\underline{w}\ .$$

From the above representation of the system, two important relationships are apparent. First the eigenvalues of the linear portion of the system are completely arbitrary, since theorems 3 and 4 showed that the eigenvalues of $\underline{A} + \underline{b}\,\underline{k}'$ and $\underline{A} - \underline{g}\,\underline{d}'$ are arbitrary. Secondly, since σ is a function only of the $\underline{x}$ state variables, the system can really be considered as two systems. One which is completely linear, has arbitrary eigenvalues, and whose state can be made to go to zero as time becomes large by proper selection of the arbitrary vector, $\underline{g}$, and a second which is just a Lur'e type problem with an additional input consisting of the state variables of the linear system, $\underline{x} - \underline{z}$, through the input matrix, $-\underline{b}\,\underline{k}'$. The differential equations describing this new Lur'e type system are

$$\dot{\underline{x}} = (\underline{A} + \underline{b}\,\underline{k}')\underline{x} + \underline{b}\,f(\sigma) - \underline{b}\,\underline{k}'(\underline{x} - \underline{z}) \tag{10}$$

The equations of the state, $\underline{x}$-$\underline{z}$, are

$$\dot{\underline{x}} - \dot{\underline{z}} = (\underline{A} - \underline{g}\,\underline{d}')(\underline{x} - \underline{z}) \tag{11}$$

and have the stable solution

$$\underline{x} - \underline{z} = e^{(\underline{A} - \underline{g}\,\underline{d}')t}(\underline{x}_o - \underline{z}_o) \tag{12}$$

Equation (10) may be rewritten in terms of (12) as

$$\dot{\underline{x}} = (\underline{A}+\underline{b}\,\underline{k}')\underline{x} + \underline{b}\,f(\sigma) - \underline{b}\,\underline{k}'\, e^{(\underline{A} - \underline{g}\,\underline{d}')t}(\underline{x}_o - \underline{z}_o) \tag{13}$$

$$= \underline{H}\,\underline{x} + \underline{b}\,f(\sigma) + \underline{g}(t)$$

where

$$\underline{g}(t) = -\underline{b}\,\underline{k}'\, e^{(\underline{A} - \underline{g}\,\underline{d}')t}\,(\underline{x}_o - \underline{z}_o) \tag{14}$$

The first two terms of (13) are just of the form of the original Lur'e problem. Aizerman and Gantmacher [2] have shown that, if this system is absolutely stable, a Lyapunov function of the form a quadratic plus the integral of the nonlinearity exists. This is represented as

$$V = L(\underline{x}) + \beta \int_o^{\sigma} f(\sigma)d\sigma \tag{15}$$

where

$$L(\underline{x}) = \underline{x}'\underline{L}\,\underline{x}$$

$$\beta = \text{constant}$$

The derivative of V along the trajectories of the Lur'e portion of (13) is

$$\dot{V} = \nabla V \cdot \dot{\underline{x}} = \sum_i [2 \sum_j l_{ij}\, x_i + \beta\, f(\sigma)c_i]\ [\sum_j a_{ij}\, x_j + b_i f(\sigma)] \tag{16}$$

It has been shown[1] that this $\dot{V}$ is negative definite and for principal cases with finite k the Lyapunov Function given by (15) is bounded by the relationships

$$L(\underline{x}) \leq V \leq L(\underline{x}) + 1/2\ \beta k\ \sigma^2 \qquad (\text{for } \beta \geq 0) \tag{17}$$

$$L(\underline{x}) + 1/2\ \beta\ k\ \sigma^2 \leq V \leq L(\underline{x}) \qquad (\text{for } \beta \leq 0)$$

Due to these inequalities, the function V is bounded from above and below by a quadratic in $\underline{x}$. Thus from (16) and (17) the ratio of $\dot{V}$ to V is bounded above by the relationship

$$\frac{\dot{V}}{V} \leq \frac{k_b ||x||^2}{-k_a ||x||^2} \quad \text{or} \quad \frac{\dot{V}}{V} \leq -\mu \tag{18}$$

where μ is a positive number.

If for (13) the Lyapunov Function given by (15) is used and $\dot{V}_a$ calculated the result is

$$\dot{V}_a = \dot{V} + \nabla V \cdot \underline{g}(t)$$

where $\dot{V}$ is given by equation (16). Dividing by V gives

$$\frac{\dot{V}_a}{V} = \frac{\dot{V}}{V} + \frac{\nabla V \cdot \underline{g}(t)}{V} \tag{19}$$

or

$$\frac{\dot{V}_a}{V} \leq -\mu + \frac{\left|\left|\sum_i 2 \sum_j l_{ij} x_i + \beta f(\sigma) c_i\right|\right| \; ||g(t)||}{||L(\underline{x}) + \beta \int_o^\sigma f(\sigma)d\sigma||}$$

From equation (12)

$$||\underline{g}(t)|| \leq M e^{-\alpha t} \tag{20}$$

where M and α are positive constants. Using this relationship the second term of (20) approaches zero as time becomes large, thus the system will be stable if the Lur'e portion of the system can be made to satisfy the Popov criterion and the eigenvalues of $\underline{A} - \underline{g}\,\underline{d}'$ are made to have negative real parts.

In order to study the stability of the system via the Popov criterion the transfer function, $G_c(s) = -\sigma(s)/f_\sigma(s)$ must be obtained. Referring to equation (10) this given by

$$G_c(s) = \frac{-\sigma(s)}{f_\sigma(s)} = -[\underline{c}', \underline{0}] \begin{bmatrix} \underline{I}\,s - \underline{A} - \underline{b}\,\underline{k}' & -\underline{b}\,\underline{k}' \\ \underline{0} & \underline{I}\,s - \underline{A} + \underline{g}\,\underline{d}' \end{bmatrix}^{-1} \begin{bmatrix} \underline{b} \\ \underline{0} \end{bmatrix} \tag{21}$$

$$= -\underline{c}'\,(\underline{I}\,s - \underline{A} - \underline{b}\,\underline{k}')^{-1}\,\underline{b}$$

which can be expanded as

$$G_c(s) = \frac{-c_n s^{n-1} - c_{n-1} s^{n-2} - \cdots - c_1}{s^n + (\alpha_1 + k_n)s^{n-1} + \cdots + \alpha_n + k_1} \tag{22}$$

where the c_i are the original numerator coefficients of the transfer function of the Lur'e system (1) and the $\alpha_i + k_i$ are completely arbitrary coefficients selected by the choice of the matrix $\underline{k}'$.

Referring to the Popov stability criterion absolute stability is insured in the sector [0,k] if any q can be found such that

$$\text{Re}\,(1 + j\,q\,w)\,G(jw) + \frac{1}{k} > 0 \tag{23}$$

If the special case where q = 0 is considered, the resulting criterion becomes

$$\text{Re}\,G(jw) + \frac{1}{k} > 0 \tag{24}$$

or

$$\text{Re}\,G(jw) > -\frac{1}{k}$$

Since in $G_c(s)$ there is always an excess of arbitrary poles over fixed zeros of at least one, (24) can always be achieved, for in fact it is always possible to place these arbitrary poles such that

$$|G_c(jw)| < \frac{1}{k} \tag{26}$$

is achieved.

The stability requirement, $|G_c(jw)| < \frac{1}{k}$, is certainly overly restrictive in many cases since, in reality, it is only necessary to make $G_c(jw)$ such that $\text{Re}(1 + jqw)\,G_c(jw) + 1/k > 0$ is satisfied. This latter requirement obviously places no restriction on the magnitude of $G_c(jw)$ in the right hand plane of its polar representation. The fact that a restrictive form of the Popov criterion can always be satisfied verifies that the estimator system can be designed which insures system absolute stability and hence the following theorem and corollary have been proven.

Theorem 5: Consider a system described by the differential equations

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,f(\sigma)$$
$$\sigma = \underline{c}'\,\underline{x} \tag{27}$$

where $f(\sigma)$ satisfies the conditions

$$f(\sigma) = 0 \qquad \text{for } \sigma = 0$$
$$0 \leq \frac{f(\sigma)}{\sigma} \leq k \qquad \text{for } \sigma \neq 0 \tag{28}$$
$$0 < k < \infty$$
$$\lim_{\sigma \to \infty} \int_0^{\sigma} f(\sigma)\, d\sigma \to \infty$$

Suppose that a completely observable output, $y_1 = \underline{d}'\,\underline{x}$, is available from the linear portion of the system, and suppose in addition, that $\underline{b}$ provides complete controllability of the linear portion of the system, i.e., the couple $(\underline{A},\underline{d}')$ is completely observable and the couple $[A,b]$ is completely controllable. Then an asymptotic estimator can be constructed whose inputs are $y_1 = \underline{d}'\underline{x}$ and $y_2 = f(\sigma)$ and whose output state is an asymptotic estimate of the state, $\underline{x}$, (i.e. $\underline{x} - \underline{z} \to 0$ as $t \to \infty$) and where the error dynamics (i.e. the eigenvalues of matrix $\underline{D}$ in $\underline{x} - \underline{z} = e^{-Dt}(\underline{x}_o - \underline{z}_o)$) are arbitrary. In addition the output of the estimator can always be used to provide feedback which will make the system absolutely stable for all $f(\sigma)$ described by (28).

Corollary 5.1. Consider a system described by (27) and (28). Suppose that the pair $[\underline{A}, \underline{c}']$ is completely observable and the pair $[\underline{A},\underline{b}]$ is completely controllable. Assume also that the only output directly available from the system is the output of the nonlinearity, $f(\sigma)$, which is a monotonic function of σ. Then since $f(\sigma)$ is monotonic the inverse nonlinear function, $f^{-1}(f(\sigma))$ can be constructed. The output of this nonlinear function is $y_1 = f^{-1}(f(\sigma)) = \sigma$. This completely observable output, $y_1 = \sigma = \underline{c}'\underline{x}$, can then be used along with, $y_2 = f(\sigma)$, as inputs to a state asymptotic estimator which will provide absolute stability of the system for all $f(\sigma)$ defined by equation (28) when used to provide feedback compensation.

Stabilization Using Only $y_1 = f(\sigma)$

The system to be investigated here is represented by the following equations

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,f(\sigma) \tag{29}$$
$$\sigma = \underline{c}'\,\underline{x}$$

with only the output

$$y_1 = f(\sigma)$$

where $[\underline{A},\underline{b}]$ is completely controllable. The output $f(\sigma)$ is assumed to be nonmonotonic so that a completely observable output of the linear portion of the system is not available.

The equations of this system and an asymptotic estimator which observes only the output $y_1 = f(\sigma)$ and provides feedback $\underline{k}'$ are:

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,\underline{k}'\,\underline{z} + \underline{b}\,f(\sigma)$$
$$\dot{\underline{z}} = \underline{A}\,\underline{z} - \underline{g}\,\underline{d}'\,\underline{z} + \underline{b}\,\underline{k}'\,\underline{z} + \underline{b}\,f(\sigma) \tag{30}$$

This can be represented from Figure 1 with y_2 removed as

$$\begin{bmatrix} \dot{\underline{x}} \\ \dot{\underline{z}} \end{bmatrix} = \begin{bmatrix} \underline{A} & \underline{b}\,\underline{k}' \\ \underline{0} & \underline{A} - \underline{g}\,\underline{d}' + \underline{b}\,\underline{k}' \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{z} \end{bmatrix} + \begin{bmatrix} \underline{b} \\ \underline{b} \end{bmatrix} f(\sigma) \tag{31}$$

Transforming the system by the transformation, $\underline{T}_1$, (7) gives

$$\begin{bmatrix} \dot{\underline{x}} \\ \dot{\underline{x}}-\dot{\underline{z}} \end{bmatrix} = \begin{bmatrix} \underline{A} + \underline{b}\,\underline{k}' & -\underline{b}\,\underline{k}' \\ \underline{g}\,\underline{d}' & \underline{A} - \underline{g}\,\underline{d}' \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{x}-\underline{z} \end{bmatrix} + \begin{bmatrix} \underline{b} \\ \underline{0} \end{bmatrix} f(\sigma) \tag{32}$$

Then for $\dot{\underline{x}} - \dot{\underline{z}}$ to be a true asymptotic estimation, $\underline{g}\,\underline{d}' = \underline{0}$ and

$$\dot{\underline{w}} = \begin{bmatrix} \underline{A} + \underline{b}\,\underline{k}' & -\underline{b}\,\underline{k}' \\ \underline{0} & \underline{A} \end{bmatrix} \underline{w} + \begin{bmatrix} \underline{b} \\ \underline{0} \end{bmatrix} f(\sigma) \tag{33}$$

The error dynamics of the linear portion of the system are, under these conditions, just those of the linear portion of the original system and stabilization is only possible under the condition that all eigenvalues of $\underline{A}$ have negative real parts, which is just the principal case. If this condition is satisfied the transfer function,

$$G_c(s) = \frac{-\sigma(s)}{f(s)} = -\underline{c}'\,(\underline{I}s - \underline{A} - \underline{b}\,\underline{k}')^{-1}\,\underline{b} \tag{34}$$

still provides completely arbitrary poles and thus the system can still be made absolutely stable over any sector [0,k]. The following theorem may now be stated.

Theorem 6: Consider a system described by the differential equations

$$\begin{aligned} \dot{\underline{x}} &= \underline{A}\,\underline{x} + \underline{b}\,f(\sigma) \\ \sigma &= \underline{c}'\,\underline{x} \end{aligned} \tag{35}$$

where $f(\sigma)$ satisfies the conditions

$$\begin{aligned} f(0) &= 0 \\ 0 \leq \frac{f(\sigma)}{\sigma} &\leq k \qquad \text{for } \sigma \neq 0 \\ 0 &< k < \infty \\ \lim_{\sigma\to\infty} \int_0^{\sigma} f(\sigma)\, d\sigma &\to \infty \end{aligned} \tag{36}$$

and whose only available output is $f(\sigma)$. Then an asymptotic estimator can be constructed whose input is $f(\sigma)$ and whose output $\underline{z}$ is an asymptotic estimate of the state, $\underline{x}$. (i.e. $\underline{z}(t) - \underline{x}(t) \to 0$ as $t \to \infty$) if the linear approximation of the system $\dot{\underline{x}} = \underline{A}\,\underline{x}$ is asymptotically stable, or in other words, a principal case. In addition, if an asymptotic estimator can be constructed, then its output can always be used to provide feedback which will insure absolute stability of the system.

The above theorem places the strict requirement that the matrix $\underline{A}$ have eigenvalues all of which have negative real parts. This requirement is a result of assuming that none of the state variables are available to be used as an input to the estimator. In many cases where no completely

observable output is available for the linear system $\dot{\underline{x}} = \underline{A}\,\underline{x}$ at least some of the states are available. If an input from this non-observable output is applied to the estimator, certain of the eigenvalues of the estimator error dynamics, $|\underline{I}s - \underline{A} - \underline{g}\,\underline{c}'|$ can be arbitrarily selected, while others are fixed. If under this condition all the eigenvalues which are fixed have negative real parts, then the system can be made absolutely stable even when the matrix $\underline{A}$ contains certain eigenvalues having non-negative real parts and a completely observable output is not available. The real requirement then is that the state variables corresponding to unstable eigenvalues of the matrix $\underline{A}$ be available as inputs to the asymptotic estimator. This is most clearly illustrated by a simple example.

Example. In the following example, a matrix, $\underline{A}$ having one unstable eigenvalue will be considered. The matrix $\underline{A} - \underline{g}\,\underline{c}'$ will be shown to be capable of stabilization for one output, $\underline{c}'_a\underline{x}$, and not capable of stabilization for another possible output, $\underline{c}'_b\underline{x}$.

Let the matrix $\underline{A}$ be given by

$$\underline{A} = \begin{bmatrix} -5 & 0 \\ 0 & +1 \end{bmatrix} \tag{37}$$

with an output

$$y = \underline{c}'_a\underline{x} = [0\ 1]\ \underline{x} \tag{38}$$

The pair $[\underline{A},\ \underline{c}'_a]$ is not completely observable since

$$|\underline{Q}| = \begin{vmatrix} \underline{c}' \\ \underline{c}'\underline{A} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 0 & 1 \end{vmatrix} = 0 \tag{39}$$

However, if the matrix $\underline{A} - \underline{g}\,\underline{c}'_a$ is considered all eigenvalues can be made stable since,

$$\underline{A} - \underline{g}\,\underline{c}'_a = \begin{bmatrix} -5 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} [0\ 1] = \begin{bmatrix} -5 & -g_1 \\ 0 & +1-g_2 \end{bmatrix} \tag{40}$$

The characteristic equation of (40) is

$$\Delta(\lambda) = (\underline{I}\lambda - \underline{A} + \underline{g}\,\underline{c}') = (\lambda + 5)\ (\lambda - 1 + g_2) \tag{41}$$

All eigenvalues of (41) have negative real parts if

$$g_2 > 1 \tag{42}$$

Thus an asymptotic estimator could be developed with

$$\underline{g} = \begin{bmatrix} 0 \\ g_2 \end{bmatrix} \tag{43}$$

where g_2 is greater than one. The system can then be made absolutely stable if $\underline{k}'$ is selected to insure that the Popov criterion is satisfied.

Consider again the same matrix $\underline{A}$, but let the output matrix be

$$\underline{c}'_b = [1\ 0] \tag{44}$$

The pair $[A,\ c'_b]$ is still not completely observable since

$$|Q| = \begin{vmatrix} \underline{c}' \\ \underline{c}'\underline{A} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -5 & 0 \end{vmatrix} = 0 \tag{45}$$

the matrix $\underline{A} - \underline{g}\,\underline{c}'$ is in this case

$$\underline{A} - \underline{g}\,\underline{c}' = \begin{bmatrix} -5 & 0 \\ 0 & +1 \end{bmatrix} - \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} [1\ 0] = \begin{bmatrix} -5-g_1 & 0 \\ -g_2 & 1 \end{bmatrix} \tag{46}$$

and the characteristic equation is

$$\Delta(\lambda) = |\underline{I}\lambda - \underline{A} + \underline{g}\,\underline{c}'| = (\lambda + 5 + g_2)\,(\lambda - 1) \tag{47}$$

which still has the unstable eigenvalue, 1, for any selection of $\underline{g}$. In this case a stable asymptotic estimator system cannot be constructed and the system is not capable of being made absolutely stable using the theory developed here.

SUMMARY

It has been shown that an asymptotic estimator can always be used to stabilize a nonlinear system which contains a true Popov type nonlinearity in the sector [0, k] where k is finite if both the nonlinear function and a completely observable output of the linear portion are available as inputs to the estimator. In the case a completely observable output is not available from the linear portion, stabilization has been shown to be possible if the original linear approximation of the system is asymptotically stable or those state variables corresponding to the unstable eigenvalues are available.

REFERENCES

1. Lur'e, A. I. Some Nonlinear Problems in the Theory of Automatic Control, Moscow, Gostekhizdat., 1951.

2. Aizerman, M. A. and Gantmacher, F. R., Absolute Stability of Regulator Systems, San Francisco: Holden Day, Inc., 1964.

3. Ogata, K. State Space Analysis of Control Systems, Englewood Cliffs, N. J., Prentice Hall, Inc., 1967.

4. Kalman, R. E., Falb, P. L., and Arbib, M. A., Topics in Mathematical System Theory New York; McGraw-Hill Book Co., 1968.

5. Luenberger, D. G. "Observing the State of a Linear System," IEEE Transactions on Automatic Control, Vol. AC-11, pp. 190-195. April, 1966.

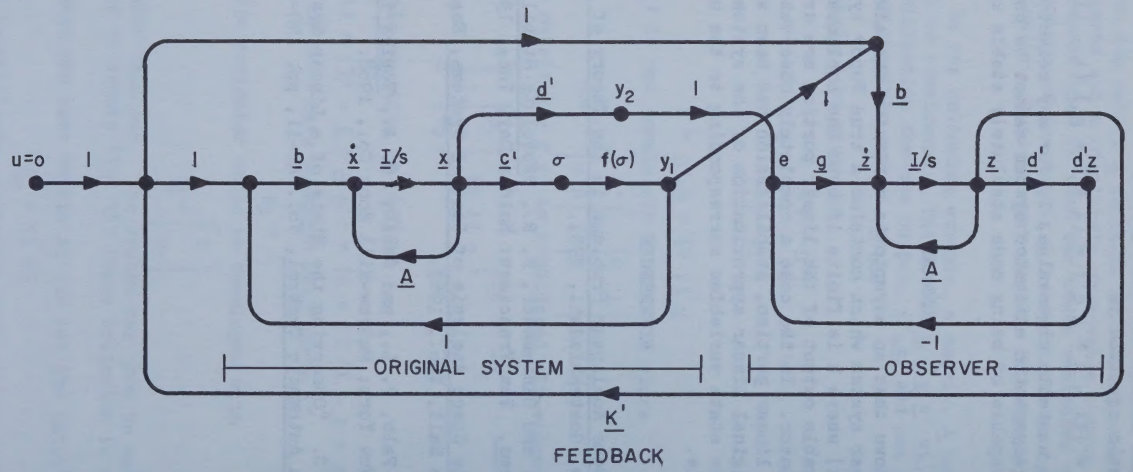


Figure 1. Signal Flow Diagram-Lur'e System and Observer for Stabilization

AN APPROACH FOR FINDING THE SINUSOIDAL STEADY STATE RESPONSE OF NONLINEAR SYSTEMS*

JOHN C. LINDENLAUB
School of Electrical Engineering, Purdue University, Lafayette, Indiana

INTRODUCTION

An important problem in the analysis of nonlinear systems in finding the steady state response due to a sinusoidal excitation. The solution of the analogous problem for linear systems is, of course, well known, and adequate techniques are available for nonlinear resistive circuits. A number of problems remain, however, when nonlinear systems with memory are considered.

The problem was motivated by the fact that circuit specifications which require distortion products to be 80 to 100 dB below the fundamental are becoming common. Such specifications are beyond the capabilities of conventional methods which solve circuit state equations by Runge-Kutta or predictor-corrector techniques. The approach outlined in this paper, unlike techniques which integrate the state equations numerically, finds the steady state response directly. This is advantageous as it avoids the cumulative numerical accuracy and stability problems which are inherent in the conventional approach [1]. Proceeding to the steady state directly also avoids the necessity of determining when the system is in steady state, i.e., when the amplitude of the transient is small enough that it may be neglected.

The technique is intended to handle circuits composed of independent current and voltage sources, resistors, capacitors, inductors, diodes and transistors. The charge control model [2] is used to represent diodes and transistors. One reason for choosing this class of circuits is that there are a number of computer programs available which formulate circuit equations from a topological description of this class of network [3,4] and it is desirable to capitalize on this existing software.

Neill [5] has approached the problem of directly obtaining the steady state solution to periodic excitations by means of a Volterra series representation of the system. His circuit model is based on that of Narayanan [9] who approximates circuit nonlinearities by a Taylor series extending to third-order terms. The method described here retains the exponential function relationships of the charge control model.

HARMONIC BALANCE - OPTIMIZATION APPROACH

The method couples the philosophy of harmonic balance with an optimization program. In the classical harmonic balance method a set of algebraic equations is obtained by equating the coefficients of the various harmonics which result from substituting a truncated Fourier series with unknown coefficients into the original equation. Unfortunately in many circumstances, the equations do not lend themselves to an analytical solution.

In the approach taken here, an optimization program is used to adjust the Fourier coefficients so as to minimize an error functional. The state equations are written in the form

---

*The work reported in this paper was carried out while the author was with the Bell Telephone Laboratories, Inc., North Andover, Massachusetts.

$$\dot{\underline{y}} = \underline{f}(\underline{y},t) \tag{1}$$

where $\underline{y}$ is the state variable vector and $\underline{f}$ is a vector of nonlinear functions. A trial solution is denoted as $\hat{\underline{y}}$ and the system error is formulated as

$$\underline{\epsilon}(t) = \dot{\hat{\underline{y}}} - \underline{f}(\hat{\underline{y}},t) \tag{2}$$

If we let $\hat{\underline{y}}$ take the form of a Fourier series with undetermined coefficients and adjust the coefficients to minimize some functional of $\underline{\epsilon}(t)$ we have, in some sense, obtained an optimum solution. Note that the system error may not have a clear physical interpretation. The condition $\underline{\epsilon}(t) = 0$ implies that the solution error, $\hat{\underline{y}}-\underline{y}$, is also zero, but in general, a nonzero $\underline{\epsilon}(t)$ does not yield any direct information concerning the solution error. This fact somewhat complicates the problem of deciding how small a system error is small enough.

Circuit state equation variables are capacitor voltages or inductor currents. If the state equations, Eq. (1), are multiplied by the capacitor or inductor (which might be a nonlinear function of the state variable) corresponding to the state variable, a Kirchhoff current or voltage law equation is obtained. The error function for this modified set of state equations, defined in a manner which is analogous to Eq. (2), has been given the name modified system error, $\underline{\epsilon}_m(t)$. Choosing the Fourier coefficients so as to minimize a functional of the modified system error has the physical interpretation of finding the solution which comes closest to satisfying the Kirchhoff law circuit equations.

The use of an optimization strategy to minimize a functional of the system or modified system error offers a numerical approach to the method of harmonic balance. The following sections discusses the various error functionals, optimization routines, and strategies that have been tried.

ERROR FUNCTIONS

The root mean square system error is perhaps the first type of error functional that comes to mind.

$$E_1 = \left(\sum_{i=1}^{N} \frac{1}{T} \int_0^T \epsilon_i^2(t)dt \right)^{1/2} \tag{3}$$

is one of four error functionals investigated. (N is the number of state variables and T is the period of the excitation.) As mentioned above, the physical significance of the system error is not as clear as that of the modified system error. The second error functional considered is the root mean square modified system error.

$$E_2 = \left(\sum_{i=1}^{N} \frac{1}{T} \int_0^T \epsilon_{m_i}^2(t)dt \right)^{1/2} \tag{4}$$

Liou [6] shows that the harmonic balance method is equivalent to choosing the Fourier series coefficients such that the set of equations given below is satisfied.

$$\int_0^T \epsilon_i(t) \cos k\omega t \, dt = 0 \qquad \begin{matrix} k = 0,1,2,\ldots M \\ i = 1,2,\ldots N \end{matrix} \tag{5a}$$

$$\int_0^T \epsilon_i(t) \sin k\omega t \, dt = 0 \qquad \begin{matrix} k = 1,2,\ldots M \\ i = 1,2,\ldots N \end{matrix} \tag{5b}$$

where $\epsilon_i(t)$ is the ith component of the system error and M is the highest harmonic in the assumed solution. Satisfying Eq. (5) is equivalent to making the solution error orthogonal to the components of the assumed solution. This technique is also known as the method of averaging.

A reasonable way to combine the requirements of Eq. (5) with an optimization program is to have the optimization program select the coefficients so as to minimize the error functional

$$E_3 = \frac{1}{T}\left\{\sum_{i=1}^{N}\sum_{k=0}^{M}\left[\left(\int_0^T \epsilon_i(t) \cos k\omega t \, dt\right)^2 + \left(\int_0^T \epsilon_i(t) \sin k\omega t \, dt\right)^2\right]\right\}^{1/2} \tag{6}$$

The fourth error functional, E_4, is obtained by applying the method of averaging to the modified system error instead of the system error. An expression for E_4 may be obtained by replacing $\epsilon_i(t)$ in Eq. (6) by $\epsilon_{m_i}(t)$.

OPTIMIZATION STRATEGY

Three optimization routines were tried in connection with this problem. The first routine used the strategies of steepest descent and least squares. The second method employed a simplex method [7] which may be described as a systematic search procedure. The third approach is based on the method of Fletcher and Powell [8]. While no serious attempt was made to compare the three programs the Fletcher-Powell program seemed to be better suited than the other two. The simplex method converged very slowly and the software surrounding the basic steepest descent-least squares program was awkward to work with.

Two strategies for combining the harmonic balance and optimization philosophies were tried. In the direct approach an M-term Fourier series is assumed at the outset and the optimization routine is used to find the best set of coefficients. When the system nonlinearities are weak the amplitudes of the higher harmonics are small compared to the fundamental. To achieve a better conditioned surface upon which to optimize, the perturbation approach was devised. The perturbation strategy first assumes that only a fundamental term is present and the optimization routine is used to determine the best set of amplitudes; next, second and third harmonic terms are added and the optimization program is called upon to find the best second and third harmonic amplitudes plus a correction term for the fundamental. Higher harmonics are added in succeeding steps with corrections being made to the lower order components. The approach is similiar to the classical perturbation method for finding harmonic solutions to nonlinear equations.

EXAMPLES

The techniques outlined above have been tried on two examples. The circuits and their corresponding state and modified state equations are given in Figs. 1 and 2 respectively. In addition the system error and

modified system error components are shown. The Kirchhoff law interpretation of the modified state equations is apparent for these examples.

Various combinations of optimization routines, error functionals and strategies were tried with the first example. A four term Fourier series approximation was used. Essentially the same result was obtained regardless of the combination used. Examination of the components of the solution vector shows that the d-c offset and fundamental terms tend to converge first, followed by the second, third and fourth harmonic terms in that order. The perturbation strategy gave error functions an order of magnitude or more smaller than the corresponding run using the direct strategy.

Because of its superior convergence properties and ease of use, only the Fletcher-Powell optimization routine was used in studying the second example. The perturbation technique gave more accurate results in the sense that with all other factors constant the resulting error functionals were 2 to 3 orders of magnitude smaller than with the direct technique. The perturbation required less computation time as well.

The second example was programmed so that it was possible to monitor the behavior of all four error functionals while optimizing with respect to one of them. It was found that the surfaces generated by the error functionals were parallel in the sense that minization of one of them tended to reduce the other error functionals as well. There was a tendency for the modified system error functionals to converge faster. This factor, coupled with their clearer physical interpretation, suggest that modified system error functionals are preferable to system error functionals.

Pairs of computer runs were made using the results of the first run as a starting point for the second. A different error functional was used on the second run to see if the solution diverged in any sense. Different starting points were also tried. No evidence of the existence of local minima was found in the examples.

CONCLUSIONS AND DISCUSSION

The basic approach--combining an optimization routine with the principle of harmonic balance--looks promising as a technique for finding the sinusoidal steady state response of nonlinear systems. The order of the system that can be handled by this approach is limited, however, by the capability of the optimization program (or computer storage). If a four harmonic Fourier series solution is assumed, nine parameters are required to specify the d-c and sine-cosine components for each state variable.

In the examples considered so far, the error function surfaces did not show any sign of having local minima. A theoretical effort to try to establish whether or not the error function surface is convex would provide valuable support for this computational approach.

Whereas the conventional state equations are more convenient for numerical integration techniques the modified state equations (along with the modified system error functionals) appear to be more appropriate for the harmonic balance approach.

REFERENCES

[1] Hildebrand, F.B., Introduction to Numerical Analysis, McGraw-Hill, 1956.

[2] Gray, P.E. et al., Physical Electronics and Circuit Models of Transistors, Vol. 2 of Semiconductor Electronics Education Committee Series, John Wiley and Sons, N.Y., 1964.

[3] Kuo, F.F. and Kaiser, J.F., System Analysis by Digital Computer, John Wiley and Sons, N.Y., 1966.

[4] Calahan, D.A., Computer Aided Network Design, McGraw-Hill, N.Y., 1968.

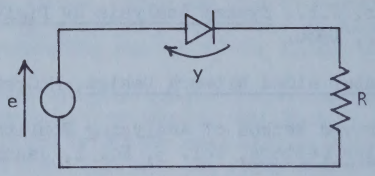
[5] Neill, T.B.M., "Improved Method of Analyzing Nonlinear Electrical Networks," Electronics Letters, Vol. 5, No. 1, January, 1969.

[6] Liou, M.L., Computational Analysis of Nonlinear Systems, Prentice Hall, (to be published), Chapter 9.

[7] Nelder, J.A. and Mead, R., "A Simplex Method for Function Minimization," The Computer Journal, Vol. 7, January, 1965, pp. 308-313.

[8] Fletcher, R. and Powell, M.J.D., "A Rapidly Convergent Descent Method for Minimization," British Computer Journal, pp. 163-168, June, 1963.

[9] Narayanan, S., "Transistor Distortion Analysis Using Volterra Series Representation," Bell Syst. Tech. J., Vol. 46, No. 5, May-June, 1967.



Circuit

State Equation $$\frac{dy}{dt} = \frac{\frac{e-y}{R} - I_0(e^{y/\lambda} - 1)}{C_j + \frac{\tau I_0}{\lambda} e^{y/\lambda}}$$

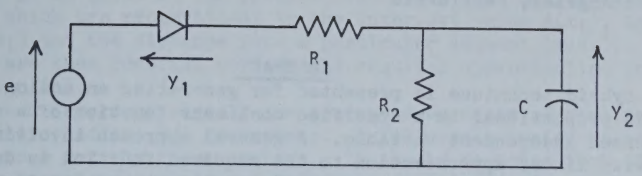
Modified State Equation $$\left(C_j + \frac{\tau I_0}{\lambda} e^{y/\lambda}\right) \frac{dy}{dt} = \frac{e-y}{R} - I_0(e^{y/\lambda} - 1)$$

Fig. 1

Single State Variable Example

I_0, λ, τ, C_j Are Parameters of The

Diode's Charge Control Model.



Circuit

$$\frac{dy_1}{dt} = \frac{\frac{e-y_1-y_2}{R_1} - I_0\left(e^{y_1/\lambda} - 1\right)}{C_j + \frac{\tau I_0}{\lambda} e^{y_1/\lambda}}$$

State Equations

$$\frac{dy_2}{dt} = \frac{e-y_1}{R_1 C} - \frac{y_2}{C}\left(\frac{R_2 + R_1}{R_1 R_2}\right)$$

$$\left(C_j + \frac{\tau I_0}{\lambda} e^{y_1/\lambda}\right)\frac{dy_1}{dt} = \frac{e-y_1-y_2}{R_1} - I_0\left(e^{y_1/\lambda} - 1\right)$$

Modified State Equations

$$C\frac{dy_2}{dt} = \frac{e-y_1}{R_1} - y_2\left(\frac{R_2 + R_1}{R_1 R_2}\right)$$

Fig. 2

Two State Variable Example

I_0, λ, τ, C_j are Parameters of the Diode's Charge Control Model.

A HYBRID APPROACH TO THE PROBLEM OF NONLINEAR FUNCTION GENERATION

R.N. NILSEN
University of California
Los Angeles, California

ABSTRACT

A hybrid technique is presented for generating an analog signal whose value is proportional to a specified nonlinear function of a digitally represented independent variable. A general approach involving the use of piece-wise linear approximation to the required function is described. An illustrative example is included which involves the generation of an analog signal proportional to the SINE of a digitally represented angle.

INTRODUCTION

Resistive ladder networks for digital-to-analog conversion (DAC) have been used extensively to generate an analog signal which is directly proportional to the coded value of a given digital register [1], [2]. A generalized synthesis procedure for resistive ladder DAC's having arbitrary, linearly weighted codes has also been developed [3]. However, very little consideration has been given to the concept of extending the resistive ladder DAC technique to provide an analog signal which is proportional to some nonlinear function of the linearly coded digital register. A method for accomplishing this nonlinear conversion is the subject of the following presentation.

The generation of an analog signal which is related to some input signal by an arbitrary function and expressed by the equation $y=f(x)$, has long been a challenging task. In an all analog environment, the linear-segment, biased-diode function generator presently has the best performance specifications [4]. In a hybrid computer environment where both analog and digital computers are available, the function generation capabilities of the digital computer in most cases preclude the necessity for requiring more than direct, linear digital-to-analog conversion. However, many special purpose hybrid (analog/digital) systems exist which require that arbitrary function generation be an integral part of the digital-to-analog conversion process. Reasons for this requirement include: (1) the lack of a fully expanded general purpose digital computer, (2) insufficient memory capability or (3) various hardware constraints such as execution speed. It was the design requirements of one of these special purpose systems which led to the development of the hybrid nonlinear conversion technique presented in this paper.

NONLINEAR D/A CONVERSION

The hybrid approach to the problem of nonlinear function generation, as presented in this paper, is based on a piecewise linear approximation to an arbitrary function. The proposed hybrid technique is similar in concept to the biased diode approach to function generation in an all analog environment. In both cases, the piece-wise linear approximation is valid only over a finite region of the independent variable and the function is assumed to be single valued within this region. The number of segments at which the slope and intercept values are to be specified is determined by the required accuracy of the analog signal and by the available resolution of the discrete independent variable.

The piece-wise linear approximation to an arbitrary function f over a finite region $[x_a, x_b]$ requires that the region of the independent variable be partitioned into n (not necessarily equal) segments. The value of n will depend on the required accuracy of the approximation. The approximation, as defined, is valid on the half open interval $[x_i, x_{i+1}]$. Over this region, the arbitrary function f is approximated by

$$f(x) = A(x_i) + B(x_i) * (x-x_i), \quad x_i \leq x < x_{i+1} \tag{1}$$

where $A(x_i) = f(x_i)$ defines the intercept and $B(x_i)=[f(x_{i+1})-f(x_i)]/[x_{i+1}-x_i]$ defines the slope of a straight line connecting the points $f(x_{i+1})$ and $f(x_i)$. For any given point x, the problem reduces to that of generating analog signals which are proportional to the intercept value $A(x_i)$, the slope value $B(x_i)$ and the distance into a particular segment $(x-x_i)$. These analog signals are then combined to form the required approximation to f as given by Equation (1).

The slope and intercept signals are generated by means of nonlinearly weighted, resistive ladder, D/A converters. A full scale analog voltage is used as a reference for these conversions. It should be noted that if a variable analog signal were used as a reference, multiplication of f by this analog signal could be accomplished without any additional hardware. The analog output of the slope matrix is then used as a reference input to a linear, resistive ladder D/A converter to provide the produce $B(x_i)*(x-x_i)$. A digital decoding network is needed to provide the required segment information used in the determination of the slope and intercept values. The most significant part of the independent variable register is used to define the particular segment to be used and the least significant portion of this same register directly defines the quantity $(x-x_i)$. A block diagram of the hybrid approach to nonlinear function generation just described is given in Figure 1.

SEGMENT DECODER

The function of the segment decoder is to provide the control necessary to select the appropriate slope and intercept information. The actual form of the decoder will depend on the particular code used to interpret the contents of the digital register. By properly partitioning the domain of the independent variable, the required segment information can be obtained from the most significant portion of the digital register. The outputs of the decoder are digital signals which control the application of the analog reference voltage to the slope and intercept networks.

The segment decoder is a multi-input, multi-output combinational network. The complexity of this combinational network is dependent on the particular code used to interpret the contents of the digital register as well as the form of the function f to be generated. A reduction in the decoder hardware complexity can be realized if symmetry conditions exist in the function f. These symmetry conditions in essence reduce the region of the independent variable over which the approximation must be implemented. For example, the SINE function over the region 2π need only be approximated over the region $\pi/2$ with proper quadrant decoding. Selection of intercept break points with reference to the particular coding of the independent variable will also reduce the decoder complexity. Selection of the number of partitions of the range $[x_a,x_b]$ to be some power of 2 will reduce the decoder complexity for binary interpretations of the digital register.

SLOPE AND INTERCEPT WEIGHTING MATRICES

The slope and intercept weighting matrices used in the nonlinear D/A conversion process are realized by use of resistive voltage divider networks of the form shown in Figure 2. The configuration of both the slope and intercept networks are the same, only the resistor values are different.

Intercept Weighting Matrix

The intercept weighting matrix is computed by assuming that the reference potential E_R is applied to the resistor corresponding to segment k; all other resistors are returned to ground (Ref. Figure 2). Equating currents at the node labeled E_k yields the following relationship

$$E_k = E_R \left[\frac{G_k}{G_o + G_T} \right] \tag{2}$$

where

$$G_T = \sum_{i=1}^{n} G_i \quad .$$

The output voltage given by Equation (2) is to be related to the required output signal as defined by the function f. The intercept voltage associated with the kth segment is proportional to the reference voltage times the function evaluated at the closed end of the segment. This relationship is expressed by the following equation:

$$E_k = C_1 E_r f(x_k) \tag{3}$$

where C_1 is a constant of proportionality. Equating Equations (2) and (3), summing both sides and solving for the constant C_1 yields

$$C_1 = \frac{G_T}{(C_o + G_T) \sum_{i=1}^{n} f(x_i)} \tag{4}$$

The value of the kth resistor in the intercept weighting matrix can now be determined. By using Equation (4) and equating Equations (2) and (3), the expression for the kth resistor is given by

$$G_k = G_T \left[\frac{f(x_k)}{\sum_{i=1}^{n} f(x_i)} \right] \text{(mhos)} \tag{5}$$

Note that the value of the kth conductance of the intercept network G_k in Equation (5) is independent of G_o, the input resistor to the scaling amplifier.

The output voltage of the intercept matrix is determined next. Using the standard transfer function for an analog feedback amplifier and Equations (3) and (4), the equation for the analog output signal is found to be

$$E_o = E_R f(x_k) \left[\frac{G_o G_T}{G(G_o + G_T) \sum_{i=1}^{n} f(x_i)} \right] \tag{6}$$

By properly choosing the feedback resistor G such that the bracketed term in Equation (6) reduces to unity, the output signal E_o is made to conform exactly to the required intercept value $f(x_k)$ times a proportionality constant E_R.

Slope Weighting Matrix

The slope resistor matrix is shown schematically in Figure 2. Required is an analog signal proportional to the slope of a line connecting intercept point (k) with intercept point (k+1). In terms of the required function f, the desired voltage at node E_k is given by

$$E_k = C_2 E_R \left[\frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} \right] \tag{7}$$

Standard forward difference operator notation yields the substitutions $\Delta x_k = (x_{k+1} - x_k)$ and $\Delta f_k = [f(x_{k+1}) - f(x_k)]$. As in the previous section, combining Equations (2) and (7), summing both sides and solving for C_2 yields

$$C_2 = \left[\frac{G_T}{G_o + G_T}\right] \sum_{i=1}^{n} \left(\frac{\Delta x_i}{\Delta f_i}\right) \tag{8}$$

By using Equation (8) and equating Equations (2) and (3), the expression for the kth resistor of the slope matrix is found to be

$$G_k = G_T \left[\frac{\Delta f_k}{\Delta x_k}\right] \sum_{i=1}^{n} \left(\frac{\Delta x_i}{\Delta f_i}\right) \text{ (mhos)} \tag{9}$$

Note that the value of the kth conductance of the slope network G_k in Equation (9) is again independent of G_o, the input resistor to the scaling amplifier.

The output voltage of the slope matrix is found by combining Equations (7) and (8) and is given by

$$E_o = E_R \frac{\Delta f_k}{\Delta x_k} \left[\frac{G_o G_T}{G(G_o + G_T)} \sum_{i=1}^{n} \left(\frac{\Delta x_i}{\Delta f_i}\right)\right] \tag{10}$$

As in Equation (6), the feed back resistor G is selected such that the bracketed term in Equation (10) is reduced to unity.

Since analog systems are fixed point in nature, some comments are in order with reference to the selection of the value of E_R and the normalization of the function f. If one selects the magnitude E_R to be the proportional to the maximum value of f over the region of interest, the values of f used in Equations (5) and (9) must be normalized to lie between zero and one. Normalization can be accomplished by replacing f by f/f_{max} in equations requiring normalization. Usually, E_R is selected to represent the maximum dynamic range of the analog system in which it is used.

LINEAR INTERPOLATION

The linear interpolation section of the hybrid function generator involves the development of an analog signal proportional to $B(x_i)(x-x_i)$. An analog signal proportional to the slope $B(x_i)$ has been generated by the nonlinearly weighted D/A converter in the previous section. Since the term $(x-x_i)$ requires an analog signal directly proportional to the difference $(x-x_i)$, a standard linearly weighted resistive D/A network can be used. To realize the product $B(x_i)(x-x_i)$, the reference signal input to this linear D/A converter is the previously generated slope signal $B(x_i)$.

The control required to gate the slope signal to the proper weighting resistors may require the decoding and/or formation of the term $(x-x_i)$ depending on the coded interpretation of the digital register. With a linearly weighted interpretation, the term $(x-x_i)$ is directly related to the less significant bits of the independent variable register. Since ladder networks for conversion of binary coded information are commercially available, one may want to first convert $(x-x_i)$ to a binary interpretation and then use a standard binary ladder network D/A converter. Again, the complexity of the linear interpolation section is directly related to the coded interpretation of the digital register.

EXAMPLE $y=\sin(\theta)$

The SINE function will be used to illustrate the design procedure for a nonlinear hybrid function generator. The angle θ is assumed to be binary encoded over the region $[0,2\pi]$. Because of symmetry conditions, the SINE function need only be approximated over the region $[0,\pi/2]$ provided proper quadrant decoding is implemented. With binary encoding, the most significant bit represents an angle of π, the next most significant bit represents $\pi/2$ and so on. For this example, the region $[0,\pi/2]$ will be partitioned into eight (8) segments. Over each segment, a piece-wise linear approximation to the SINE function will be used. It should be noted that by merely

reordering the decoder inputs, the COSINE function could be realized using the same resistor network values.

The segment decoder consists of two parts; a quadrant decoder and a slope/intercept decoder. The quadrant decoder consists of the 180° and 90° bits of the angle register. The 180° bit is used to determine the polarity of the reference voltage in order to insure the proper output sign. The 90° bit is used to select and properly decode all lessor significant bits. The course partitioning of the region $[0-\pi/2]$ into eight segments requires only three bits, namely 45°, 22.5° and 11.25°, be interrogated. These three bits and their complements plus quadrant information are then decoded via a combinational network to provide control signals for each of the following intercept points: 0°, 11.25°, 22.50°, 33.75°, 45°, 56.25°, 67.50°, 78.75°.

The eight equations defining the requirements of the slope and intercept control lines are derived from the bit configurations of the binary coded angle register. The eight slope/intercept control lines can be formed using 14 "AND" gates and 3 inverters. The decoding is accomplished by pairwise forming the exclusive "OR" of the 90° bit with each of the three lessor order bits, and appropriate gating.

Having developed the control signals for the slope and intercept matrices, the values of the slope and intercept resistor networks are then evaluated. Using Equations (5) and (9), the values of the intercept and slope resistors are computed and presented in Tables I and II. The free parameter involved in the design of these networks is G_T. Factors influencing the selection of G_T include: power requirements and system bandwidth. A small value for G_T will minimize the power requirements of the system but tends to decrease the system bandwidth whereas larger values of G_T will increase bandwidth at the expense of power consumption. As a compromise, a R_T value of 2K ohms was selected for the intercept matrix and 100K ohms for the slope matrix. It should be noted that for the particular type of D/A conversion network used, a wide range of resistor values can be expected.

Linear interpolation required to generate $B(\theta_i)(\theta-\theta_i)$ requires a standard linear D/A converter with binary inputs and a decoder to control the converter inputs. Symmetry properties of the SINE function require a reflection of the piece-wise linear approximation about the $\pi/2$ and $3\pi/2$ break-points. The control for the term $(\theta-\theta_i)$ can be derived by again taking the exclusive "OR" of each bit less than 11.25° with the 90° bit. Care should be taken to insure proper phase relationships. The product of $B(\theta_i)$ $(\theta-\theta_i)$ is formed by applying the slope voltage $B(\theta_i)$, formed previously, as the reference input to the linear D/A converter. The sign of the interpolation term is determined by the sign of $B(\theta_i)$.

Finally, an operational amplifier is used to combine the intercept voltage with the linear interpolation voltage to form the desired result. A saving of one operational amplifier can be realized by using either the summing amplifier of the intercept matrix or that of linear D/A converter as also the final summing amplifier.

ERROR SOURCES

A generalized analytic approach to the error analysis of the hybrid nonlinear function generator is beyond the scope of this paper. However, some comments should be made as to the sources and relative importance of some major contributors to the total system error.

In designing any particular nonlinear function generator, a certain upper bound is usually placed on the tolerable or allowable system error. This upper bound directly controls such system parameters as register word-length, conversion speed, and the approximation error.

The coarseness of the grid (i.e., the number of segments) governs the approximation error and is directly controlled by the upper bound on allowable error. A compromise must be made between too many segments which minimize approximation error but are costly in hardware and too few segments

which minimize hardware but may contribute a dominate error factor to the overall system. A rule of thumb is to constrain approximation errors to be less than 10% of the total allowable error leaving a wide margin for the cumulative effects of other error sources.

Quantization error is another major error source. The resolution of the digital register must be better than the required resolution of the output. For example, analog accuracies of $1:10^4$ require at least a 14 bit representation of the binary coded information. Again, a cost factor forces the selection of the shortest word length commensurate with the system error bound.

In the conversion process itself several factors contribute to the total error. A dominant factor is the tolerance of various physical components and in particular the tolerances of the resistors used in the various networks. Other factors which contribute errors to a lesser degree are sources such as amplifier drift, reference voltage errors, and any off-set voltages in the analog switches.

Bounds on individual error sources can be made for any given realization. Certain error sources are minimized by proper choice of design parameters, others are found to be less controllable. Experience has shown that $1:10^3$ is not an unreasonable overall error specification for this nonlinear function generator and in particular cases $1:10^4$ can be realized.

CONCLUSIONS

This paper has presented a hybrid approach to the problem of nonlinear function generation. This technique is particularly useful in the area of special purpose hybrid systems such as process control, and weapon control systems. This technique can also be incorporated as part of an implicit function generation scheme to generate the inverse of any nonlinear function $x=f^{-1}(y)$ and was actually used to provide cartesian-to-polar coordinate conversion.

The advent of high speed analog switches and deposited precision resistive arrays has enhanced the potential of this approach through increased speed and reduced size.

ACKNOWLEDGMENT

The author wishes to acknowledge the contributions and many helpful comments of Mr. W. Janes and Mr. S.I. Finkel of Vitro Laboratories, West Orange, New Jersey during the design and development of the SINE/COSINE D/A converters.

REFERENCES

[1] Suskind, A.K. Ed., Notes on Analog to Digital Conversion Techniques, Technology Press (MIT) 1957, pp. 5-29 to 5-54.

[2] Hoeschele, D.F. Jr., Analog-to-Digital/Digital-to-Analog Conversion Techniques, John Wiley & Sons, Inc., 1968, pp. 144-194.

[3] Aaron, M.R., Mitra, S.K., "Synthesis of Resistive Digital-to-Analog Conversion Ladders for Arbitrary Codes with Fixed Positive Weights," IEEE Transaction on Electronic Computers, V. EC-16, N. 3, pp. 277-281, June 1967.

[4] Korn, G.A., Electronic Analog and Hybrid Computers, McGraw-Hill, Inc. 1964.

[5] Schmid, H., "Linear-Segment Function Generator," IRE Transaction on Electronic Computers, V. EC-11, N. 12, pp. 780-788, Dec. 1962.

[6] Chapelle, W.E., "Hybrid Techniques for Analog Function Generation," Proc. Spring Joint Computer Conf., 1963, p. 213.

TABLE — I

INTERCEPT MATRIX

G-TOTAL = 5×10^{-4} (mhos) $\sum_{k=1}^{8} \sin(\theta_k) = 4.576585$

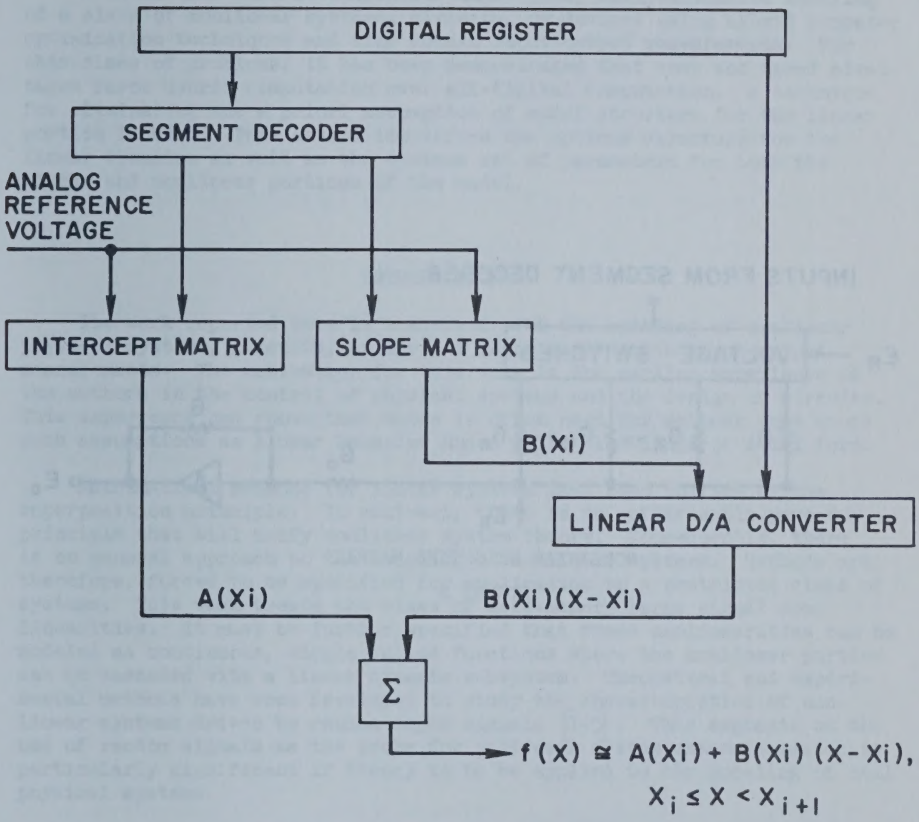
k	θ_k	$G_k \times 10^{-6}$ (mhos)	$R_k \times 10^{3}$ (ohms)
1	0°	0.000	inf.
2	11.25°	21.319	46.908
3	22.50°	41.809	23.919
4	33.75°	60.697	16.475
5	45.00°	77.253	12.945
6	56.25°	90.840	11.008
7	67.50°	100.94	9.9073
8	78.75°	107.15	9.3325

TABLE — II

SLOPE MATRIX

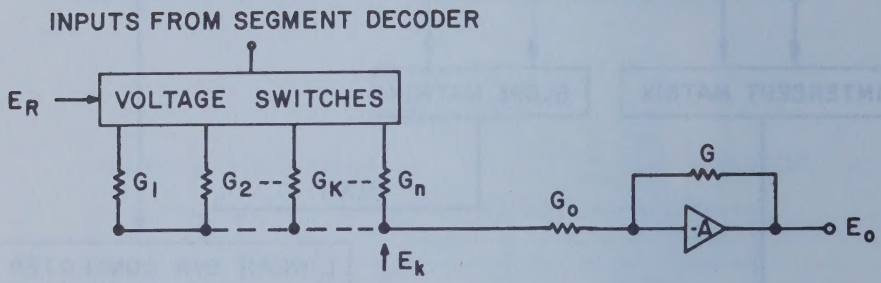
G-TOTAL = $1. \times 10^{-5}$ (mhos) $\sum_{k=i}^{8} \left(\frac{\Delta\theta_k}{\Delta f_k} \right) = 21.85727$

k	$G_k \times 10^{-6}$ (mhos)	$R_k \times 10^{3}$ (ohms)
1	217.17	4.6047
2	208.83	4.7887
3	192.45	5.1960
4	168.69	5.9281
5	138.44	7.2234
6	102.87	9.7211
7	63.346	15.786
8	21.389	46.752



NONLINEAR D/A CONVERTER — BLOCK DIAGRAM

FIGURE 1



NONLINEAR WEIGHTING MATRIX

FIGURE 2

DIRECT MODELING OF NONLINEAR SYSTEMS USING HYBRID COMPUTER OPTIMIZATION TECHNIQUES*

CHARLES H. BECK, ROBERT L. DRAKE, AND MING H. KUO
Department of Electrical Engineering
Tulane University, New Orleans, Louisiana

ABSTRACT

A method has been developed for high-speed, fully-automated modeling of a class of nonlinear systems, circuits, or devices using hybrid computer optimization techniques and time domain input-output measurements. For this class of problems, it has been demonstrated that cost and speed advantages favor hybrid computation over all-digital computation. A technique for eliminating the a priori assumption of model structure for the linear portion is used. This method identifies the optimum structure for the linear dynamics as well as the optimum set of parameters for both the linear and nonlinear portions of the model.

INTRODUCTION

The work reported here is concerned with the modeling of nonlinear physical systems, circuits, or devices using time-domain input-output measurements. The motivation for this work is the earlier experience of the authors in the control of physical systems and the design of circuits. This experience has shown that there is often need for methods that avoid such assumptions as linear behavior and a priori knowledge of model form.

Mathematical methods for linear systems have been unified by the superposition principle. In contrast, there is no sufficiently powerful principle that will unify nonlinear system theory. Consequently, there is no general approach to the modeling of nonlinear systems. Methods are, therefore, forced to be specified for application to a restricted class of systems. This work treats the class of zero-memory large signal nonlinearities. It must be further specified that these nonlinearities can be modeled as continuous, single-valued functions where the nonlinear portion can be cascaded with a linear dynamic subsystem. Theoretical and experimental methods have been developed to study the characteristics of nonlinear systems driven by random input signals [1-5]. This emphasis on the use of random signals as the probe for nonlinear system characteristics is particularly significant if theory is to be applied to the modeling of real physical systems.

A method has been developed for high-speed, fully-automated modeling of a class of nonlinear systems driven by ergodic band-limited white random inputs. A situation frequently encountered is that not only are some or all of the parameters of a model unknown, but the structure of the model may also be unknown. A technique has also been developed for eliminating the usual assumption of model structure for the linear dynamic subsystem with the result that the optimum structure, as well as the optimum set of parameters, is identified [6]. For this class of problems, it has been demonstrated that cost and speed advantages favor hybrid computation over all-digital computation [7].

*This research is partially supported by contracts NAS12-655 and NAS12-2059.

MATHEMATICAL MODEL

In many actual physical systems, the nonlinearities occur primarily in the large signal behavior whereas a linear idealization leads to a very reasonable approximation of small signal system operation. The class of systems under consideration is described as a zero-memory, subsystem [8] as shown in Figure 1.

For a single input--single output nonlinear element, the output m is approximately equal to the output of a piecewise linear representation for a given input r [9]. The output of a zero-memory, single valued nonlinear element can be expressed as

$$m = F\left[r, m_o, \alpha_1, \cdots \alpha_n\right] \tag{1}$$

where each parameter α_i is the slope of a segment of the piecewise linear function.

For a bounded input, quantization of the input dynamic range yields

$$\Delta r \triangleq \frac{R_{max} - R_{min}}{n} \tag{2}$$

where R max and R min are upper and lower bounds on the input, and n is the number of quantization segments. To specify the parameters of the nonlinear element, let

$$\alpha_i \triangleq \frac{dm}{dr}\Big|_r \qquad i\Delta r < r \leq (i+1)\Delta r \tag{3}$$

$$1 \leq i \leq n$$

For simplicity, let

$$m_o = m\,(R_{min}) \tag{4}$$

and

$$m_i = m\,(i\,\Delta\,r) \triangleq m\,(r_i) \tag{5}$$

Then

$$m = \sum_{i=1}^{n} \left(\left[r-r_i\right]\,\alpha_i + m_{i-1}\right)\left[u\,(r-r_i) - u(r-r_{i+1})\right] \tag{6}$$

Since m_{i-1} can be expressed as

$$m_{i-1} = \Delta\, r \sum_{k=1}^{i-1} \alpha_k + m_o \tag{7}$$

as indicated in Figure 2, Equation (5) becomes

$$m = \sum_{i=1}^{n} \left[(r-r_i)\,\alpha_i + \Delta\, r \sum_{k=1}^{i-1} \alpha_k + m_o\right]\left[u(r-r_i) - u(r-r_{i+1})\right] \tag{8}$$

Therefore, the output of the nonlinear element is

$$m = G\left[r,\ n,\ m_o,\ \alpha_1,\ \cdots,\ \alpha_n\right] \tag{9}$$

This nonlinear portion is cascaded with a typical linear subsystem described by the state differential equation

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}\,\underline{M} \tag{10}$$

If the linear portion is a network of known form it will be desirable to use a transform approach for representation of the linear dynamics. This can be accomplished easily by obtaining the transfer function using the computer-aided network analysis program NASAP* [10].

The dynamic output of the total single input—single output cascaded system is

$$C_m(t) = f\left[r(t),\ \Delta r,\ m_o,\ \underline{\alpha}_i,\ \underline{a}_{jk}\right] \qquad 1 \le i \le n \tag{11}$$

where the $\underline{a}_{jk}$ are related to parameters of the linear subsystem.

MODELING METHOD

The modeling method makes use of hybrid computation and parameter optimization techniques [11]. The system block diagram of the hybrid computer which the authors have developed is shown in Figure 3. The use of the hybrid computer is indicated by the block diagram of Figure 4. An assumed nonlinear model structure is programmed on the analog portion of the hybrid computer. The automated patching of the linear portion is controlled by the structure optimization algorithm portion of the digital program. The parameters of both the linear and nonlinear portions of the model are set under control of the digital program. High speed digital parameter units (DPU) have been constructed to provide for this automated feature.

The transient response of the model for a given input function is obtained for an initial set of parameter values, and the performance index (IP) is computed. Then, the response is again obtained for a variation of parameter values without operator intervention. The performance index from this second run is compared with that from the first run. An optimization algorithm is used to check for improvement in the performance index. Based on this decision, a search algorithm is used to select a new set of parameters. This process is repeated until the optimum set of parameters is identified for the assumed model structure. The form of the model structure of the linear portion is changed to the next higher order. Then, the optimum set of parameters is obtained for the second model structure, and the optimum performance index is compared with that for the former model structure. This process is continued under control of the optimization algorithm until the optimum model structure, as well as the optimum set of parameters for this model, is identified. Parameters of the nonlinear portion are included by simply extending the parameter vector.

*Network Analysis for System Applications Program developed by NASA/Electronics Research Center

It has been found that the standard canonical forms for linear dynamic models should be modified for more efficient hybrid optimization in "growing" models. The growth model structure shown in Figure 5 has been devised in this work to provide better initial parameter estimates than obtained with the canonical forms [6]. Also, the factored form of the transfer function is always available with the growth model structures. Figure 5 illustrates correct identification of a linear second-order response with a second-order model. If zeros are present, they are included by the identification of additional finite parameters.

PERFORMANCE INDEX

In obtaining the "best" values for the parameters of the model, it is necessary to establish a comparison criterion or performance index. An obvious choice is an index based on the difference (error) between the actual transient response of the physical system being modeled and the transient response of the computer model operated in parallel with and driven by the same input as the physical system [12]. The choice of performance index will be influenced by the type and the characteristics of the driving function.

In some cases, the economic or safety considerations associated with experiments which disturb the system or drive it outside of the normal operating range may preclude the use of classical analysis techniques. Only normal operating data with its inherent noise components is available in these cases. To assure reliability when system parameters are variable, on-line continuous model updating may be required. If the system is high-order or nonlinear, the solution time for all-digital computation could be excessive. If it can be assumed that parameters are fixed, then it is possible to record driving and response functions for off-line modeling [13]. In other situations, the normal system operation can be shut down and off-line excitation is possible. It is then possible to drive the system with a random noise signal or a step function input.

If the system can be assumed stationary and ergodic for any of the situations in which a random driving function is used, then ensemble average can be replaced by time averages. In such cases the mean value may be estimated from the finite time average

$$\overline{X(t)}_T = \frac{1}{T}\int_0^T x_j(t)\,dt \qquad (12)$$

where $x_j(t)$ is a particular sample function of the process, and T large compared to the lower bound of the frequency spectrum $X(\omega)$. With the presence of low frequency signal components, excessively long integration times are required to reach the mean value. A mean square error criterion is commonly used as a measure of the performance index for the study of random process. This function can be expressed as

$$\overline{IP}_T = \frac{1}{T}\int_0^T e^2(t)\,dt \qquad \text{for T long} \qquad (13)$$

Since the error is the difference between the actual transient response of the physical system and the transient response of the computer model, Equation (13) becomes

$$\overline{IP}_T = \frac{1}{T}\int_0^T \left[C(t) - C_m(t)\right]^2 dt \qquad (14)$$

Substituting for $C_m(t)$ given in Equation (11) there results

$$\overline{IP}_T = \frac{1}{T}\int_0^T \left\{C(t) - f\left[r(t), m_o, \underline{\alpha}_i, \underline{a}_{jk}\right]\right\}^2 dt \qquad (15)$$

Hence, the optimum or minimum value of IP is a function of the parameters of the nonlinear portion of the model as well as those of the linear subsystem.

PRACTICAL IMPLEMENTATION AND EXPERIMENTAL RESULTS

The experimental determination of the characteristics of a given nonlinear system, based upon the described method, is demonstrated as shown in Figure 4. The given nonlinear system is characterized by the expression

$$\dot{C}(t) + 3.14\, C(t) = 2\, m(t) \qquad (16)$$

where m(t) is the characteristic of a nonlinear element expressed by

$$\begin{aligned} m(t) &= 0 && \text{for } 0 > r(t) \\ &= 2r(t) && \text{for } 0 \leq r(t) \leq 1 \\ &= 2 && \text{for } r(t) > 1 \end{aligned} \qquad (17)$$

The given system and the modeled system, which is constructed of a cascade combination of a nonlinear element and a linear dynamic subsystem, are both simulated on the analog computer with random input. Since an entire ensemble of driving functions may arise under actual operating conditions, a random driving function such as an ergodic, band-limited, white noise generator with Gaussian amplitude distribution can appropriately be used for excitation. For practical reasons, the mean square error performance index given in Equation (13) has been modified to give a weighted integral square error (WISE) function of the form

$$IP = k\int_0^T e^2(t)\, dt \qquad (18)$$

Observation time, T, for the measurement of the performance index is determined by a preliminary parameter perturbation procedure. T is acceptable if the variance of the measurement of the performance index, defined by WISE, is small compared with the difference in absolute values of the IP measurements. Figure 6 shows the relation between T and the performance index measurements with different parameter perturbations. Obviously, from Figure 6, if an arbitrary point in the parameter space is far away from the desired optimum point, the error signal will contain a relatively large mean value; also, the absolute value of the variance of PI measurements will be large. Similarly, small absolute variance of PI measurements will occur for small parameter perturbations.

In this hybrid optimization algorithm, a large parameter increment step $\Delta\underline{p}$ is desired initially to prevent the measurements of PI from overlapping such that the difference of PI measurements with different parameter sets will be distinguishable. For the point close to the optimum point in parameter space, a small step increment of p is acceptable

because of the small absolute variance of the PI measurements.

For a given system as specified in Equation (17), the optimum structure was found to be first order and the parameters of the linear subsystem were found to be 1.89 for K_8 and 3.11 for a. The nonlinear characteristics of the optimum model was identified as:

$$\begin{aligned} m(r) &= 0 && \text{for } 0 \geq r \\ &= 1.86\, r && \text{for } 0 \leq r \leq 1.1 \\ &= 2.05 && r > 1.1 \end{aligned} \qquad (19)$$

The off-line example reported here assumes partitioning of the large signal nonlinearity and the linear dynamics. The optimum parameter set for the nonlinear portion can be identified independently of the optimum model structure and parameter set for the linear subsystem. A small step input is used to probe for the linear subsystem characteristics. Then, a random noise source is used to probe independently for the characteristics of the nonlinear portion.

CONCLUSIONS AND EXTENSIONS

This modeling method accomplishes on-line or off-line high-speed fully automated identification of nonlinear dynamic systems without a priori knowledge of the structure of the linear portion. In addition, this method makes immediately available a physically realizable direct model constructed of analog computer elements. The experience of the authors to date indicates at least an order of magnitude ratio in speed and in capital expenditure, both favoring hybrid computation over all-digital computation, for low order systems of the class described. As the order of the linear dynamics increases, the favorable speed ratio is further increased. Preliminary studies have indicated no additional difficulties with extending this method to include small-signal nonlinearities or to combinations except that the length of the parameter vector is increased. This work has also incorporated some investigations of the use of this modeling method to automatically obtain noninteracting control of multivariable nonlinear systems. These extensions will be reported elsewhere. The present work of the authors is directed to developing original hardware to permit structure optimization of the nonlinear portion of the model.

REFERENCES

1. Booton, R. C., "Nonlinear control systems with random inputs," IRE Trans. on Circuit Theory, vol. CT-1, pp. 9-18, March 1954.
2. Chuang, K. and L. F. Kazda, "A study of nonlinear systems with random inputs," Trans. AIEE, Applications and Industry, vol. 78, part II, pp. 100-105, 1959.
3. Ku, Y. H., "On nonlinear networks with random inputs," IRE Trans, on Circuit Theory, vol. CT-7, pp. 479-490, December 1960.
4. Pervozvanski, A. A. Random Processes in Nonlinear Control Systems, Academic Press, New York, 1965.
5. Price, R., "A useful theorem for nonlinear devices having Gaussian inputs," IRE Trans. on Information Theory, vol. IT-4, pp. 69-72, June 1958.
6. Beck, C. H., R. L. Drake, and M. H. Kuo, "System modeling and structure optimization using hybrid computer techniques," Proceedings, Cornell Conference on Computerized Electronics, Ithaca, N. Y., August 26-28, 1969.

7. Beck, C. H., R. L. Drake, and M. H. Kuo, "A NASAP module for direct design of linear dynamic circuits using a hybrid computer optimization technique," Conference Record, Twelfth Midwest Symposium on Circuit Theory, pp. VIII 2.1-VIII 2.9, Austin, Texas, April 1969.
8. Narendra, K. S. and P. G. Gallman, "An iterative method for the identification of nonlinear systems using a Hammerstein model," IEEE Trans. on Automatic Control, vol. AC-11, pp. 546-550, 1966.
9. Lion, P. M. "Rapid identification of linear and nonlinear systems," JACC, pp. 605-615, August 1966.
10. Happ, W. W., "Network analysis for systems applications program: present capabilities of a maintained program" IEEE Int. Conv. Record, pt. 5, pp. 64-88, March 1967.
11. Beck, C. H. and R. L. Drake, "A hybrid-NASAP module for direct design of linear dynamic circuits," IEEE Trans. on Electronic Computers, vol. EC-12, pp. 31-35, September 1969.
12. Hsia, T. C. and V. Vimolranich, "An on-line technique for system identification," IEEE Trans. on Automatic Control, vol. AC-14 no. 1, pp. 92-96, February 1969.
13. Kirvartis, K. and K. S. Fu, "Identification of nonlinear systems by stochastic approximation," JACC, pp. 255-264, August 1966.

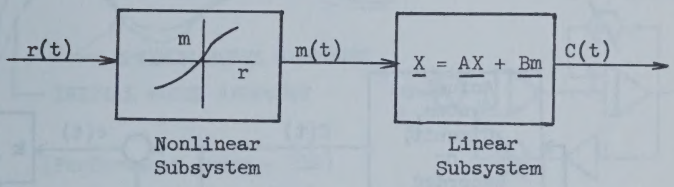


Figure 1. Block Diagram for Nonlinear System Model

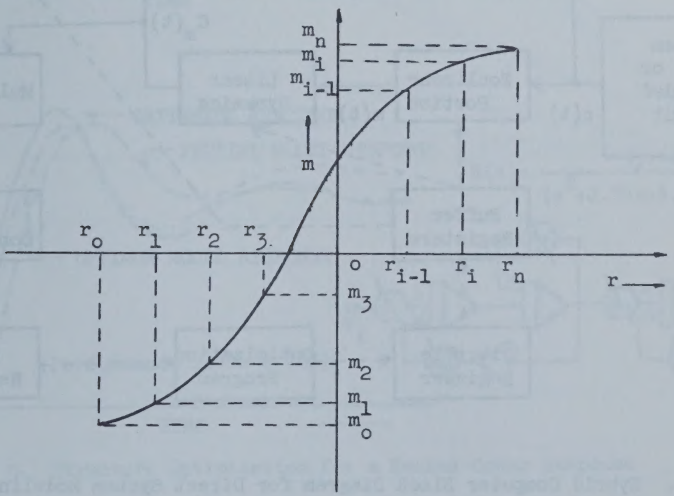


Figure 2. Single Valued Nonlinear System Model Characteristics

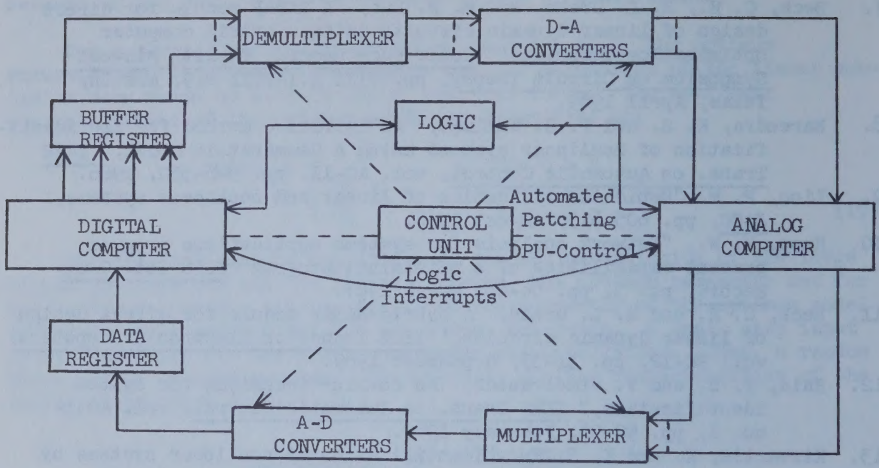


Figure 3. Tulane Hybrid Computer System Diagram

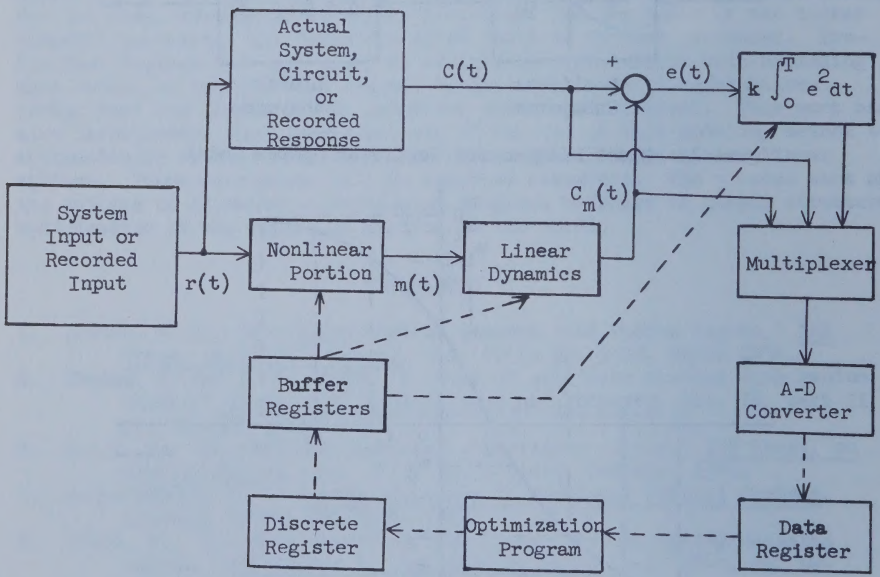


Figure 4. Hybrid Computer Block Diagram for Direct System Modeling

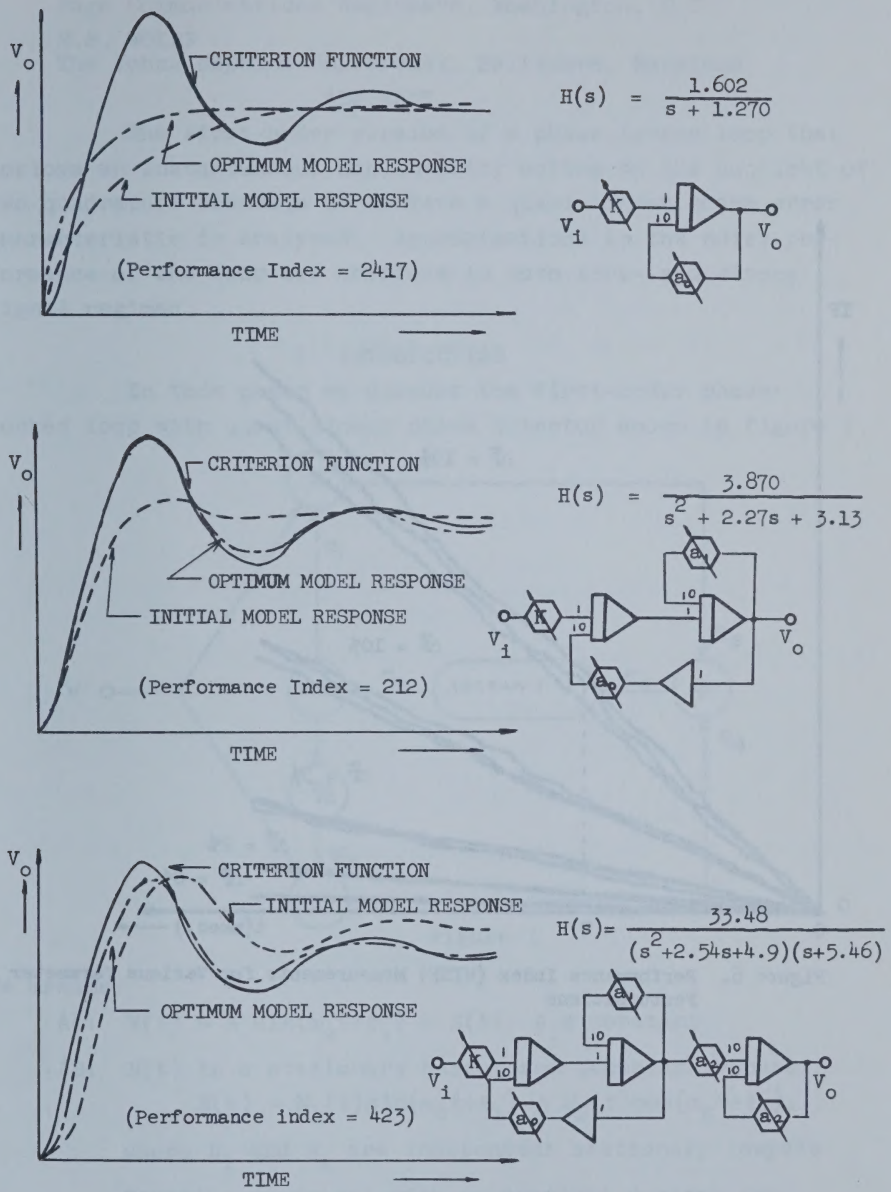


Figure 5. Structure Optimization for a Second-Order Response

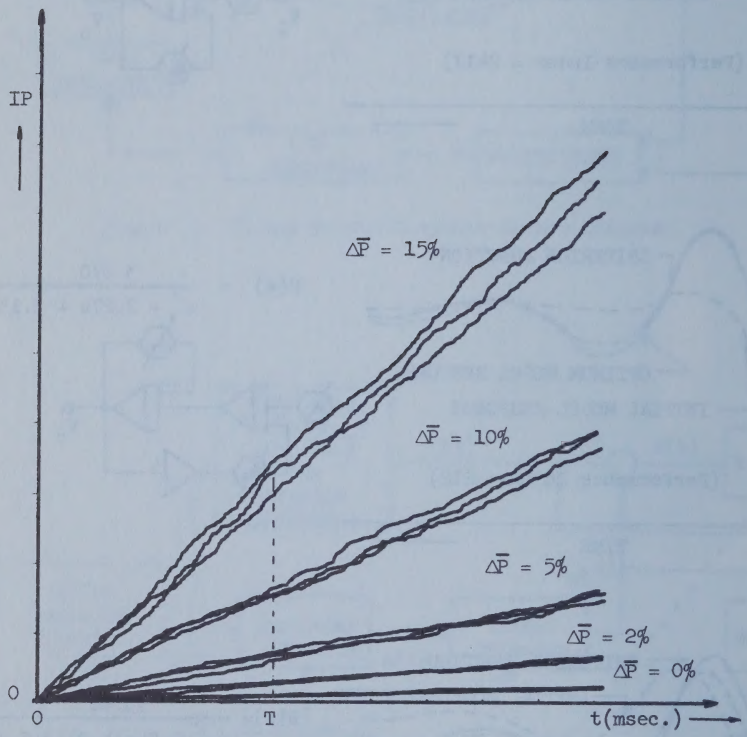


Figure 6. Performance Index (WISE) Measurements for Various Parameter Perturbations

A PHASE-LOCKED LOOP WITH QUASI-LINEAR PHASE DETECTOR

J.H. YANG
Page Communications Engineers, Washington, D.C.
S.S. WOLFF
The Johns Hopkins University, Baltimore, Maryland

ABSTRACT

The first-order version of a phase-locked loop that employs an instantaneous nonlinearity acting on the quotient of two quadrature channels to achieve a quasi-linear phase error characteristic is analyzed. Approximations to the noisy performance of the loop are obtained in both zero- and strong-signal regimes.

1. INTRODUCTION

In this paper we discuss the first-order phase-locked loop with quasi-linear phase detector shown in figure 1.

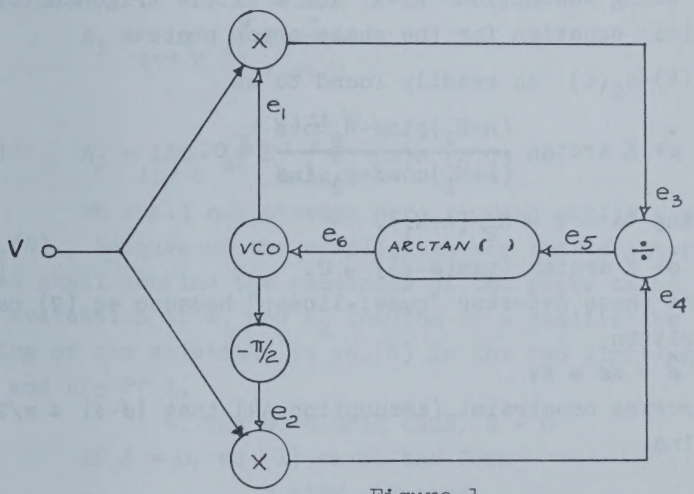


Figure 1

We assume:

A1: $V(t) = A \sin(\omega_c t+\theta_1) + N(t)$, θ_1 a constant.

A2: $N(t)$ is a stationary narrowband Gaussian process

$$N(t) = N_1(t)\sin(\omega_c t+\theta_1) + N_2(t)\cos(\omega_c t+\theta_1),$$

where N_1 and N_2 are independent stationary lowpass Gaussian processes with no spectral content for $|\omega| > \omega_c/2$, each with mean zero and covariance function $R(s,t) = \sigma^2\rho(\tau)$ $(\tau = s-t)$.

A3: The symbol $\div$ stands for an idealized analogue divider: $e_5(t) = e_3(t)/e_4(t)$.

A4: The symbol arctan( ) stands for an instantaneous nonlinearity with input-output relation $e_6(t) = \arctan e_5(t)$, $|e_6(t)| \leq \pi/2$.

A5: The ideal voltage-controlled oscillator (VCO) is described by the pair of equations
$e_1(t) = 2\cos(\omega_c t+\theta_2)$, $\dot{\theta}_2(t) = Ke_6(t)$,
K being termed the "VCO gain"

A6: The symbol $-\pi/2$ stands for a perfect 90° phase-shifter, so that $e_2(t) = 2\sin(\omega_c t+\theta_2)$.

A7: The symbols $\otimes$ represent ideal low-pass multipliers.

Using assumptions A1-A7 and a little trigonometry, the dynamical equation for the phase error process $\phi(t) \triangleq \theta_1(t)-\theta_2(t)$ is readily found to be

$$\dot{\phi} + K \arctan \frac{(A+N_1)\sin\phi - N_2\cos\phi}{(A+N_1)\cos\phi + N_2\sin\phi} = 0. \qquad (1)$$

or, defining $\tan\psi = N_2/(A+N_1)$,

$$\dot{\phi} + K \arctan(\tan(\phi-\psi)) = 0. \qquad (2)$$

We term the phase detector "quasi-linear" because eq.(2) can only be written

$$\dot{\phi} + K\phi = K\psi$$

if the hardware constraint (assumption A4) that $|\phi-\psi| \leq \pi/2$ is kept in mind.

2. THE NOISE-FREE CASE

If $N(t) \equiv 0$, eq. (2) can be written

$$\dot{\phi} + K\phi^* = 0, \qquad (3)$$

where ϕ^* is the modulo-π version of ϕ that lies in the interval $(-\pi/2,\pi/2)$. Given an initial phase error $\phi(0)$, it is immediate that the solution of eq.(3) is an exponential with time constant $1/K$ asymptotic to that multiple of π closest to $\phi(0)$.

We note in passing that our motivation to study the system of figure 1 was that eq.(3) shows the phase error dynamics to be independent of the input signal strength A.

3. THE FOKKER-PLANCK EQUATION FOR THE STEADY-STATE PHASE ERROR DISTRIBUTION

When N(t) is not identically zero, $\phi(t)$ is a random process; from time to time the loop skips cycles (or in this case, half-cycles) just as other phase-locked loops do, and the probability distribution of ϕ spreads out indefinitely so that, properly speaking, no steady state is ever reached. Mimicing the methods of Viterbi (1963) and Tikhonov (1959), however, we shall consider ϕ modulo π, which does have a limiting probability density $p(\phi)$ that can be found in principle by solving the steady-state version of the Fokker-Planck equation

$$A_1 p - \frac{A_2}{2}\frac{dp}{d\phi} = C \tag{4}$$

in which the parameters A_1 and A_2 are computed (using the dynamical equation (1)) from the relations

$$A_1 = \lim_{\Delta t \to 0} \frac{K}{\Delta t} E\int_{t-\Delta t/2}^{t+\Delta t/2} \dot{\phi}(s)ds \tag{5}$$

$$A_2 = \lim_{\Delta t \to 0} \frac{K^2}{\Delta t} E\int_{t-\Delta t/2}^{t+\Delta t/2}\int \dot{\phi}(s)\dot{\phi}(u)ds\,du. \tag{6}$$

We shall not attempt here to find explicit solutions to eq.(4). Because of the complicated way that ϕ enters eq. (1), we shall confine the remainder of the paper to an approximate evaluation of A_1 and A_2 leading to a qualitative understanding of the solutions to eq.(4) in the two limiting cases $A = 0$ and $A/\sigma >> 1$.

4. THE NO-SIGNAL CASE, A = 0

If $A = 0$, eq.(1) takes the form

$$\dot{\phi} + K \arctan \frac{N_1 \sin\phi - N_2 \cos\phi}{N_1 \cos\phi + N_2 \sin\phi} = 0. \tag{7}$$

Using assumption A2, it is easily found that $X = N_1\cos\phi + N_2\sin\phi$ and $Y = N_1\sin\phi - N_2\cos\phi$ are (for fixed ϕ) independent stationary Gaussian processes, each with zero mean and covariance function $\sigma^2\rho(\tau)$. Taking the expectation inside the integral, eq.(5) becomes

$$A_1 = -K \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} ds \iint_{-\infty}^{\infty} \arctan \frac{y}{x} \, e^{-\frac{(x^2+y^2)}{2\sigma^2}} \frac{dydx}{2\pi\sigma^2},$$

which yields, upon introduction of polar coordinates appropriate to assumption A4,

$$A_1 = -K \int_{-\pi/2}^{\pi/2} \theta d\theta \int_{-\infty}^{\infty} |r| e^{-r^2/2\sigma^2} \frac{dr}{2\pi\sigma^2} = 0. \qquad (8)$$

Similarly, eq.(6) for A_2 becomes

$$A_2 = K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \int ds du \int_{-\pi/2}^{\pi/2} \int \theta_s \theta_u d\theta_s d\theta_u$$

$$\times \int_{\infty}^{\infty} \int |r_s r_u| \frac{dr_s dr_u}{4\pi^2\sigma^4(1-\rho^2)} \, e^{-\frac{r_s^2+r_u^2-\rho r_s r_u \cos(\theta_s \theta_u)}{2\sigma^2(1-\rho^2)}} \qquad (9)$$

where $\rho = \rho(s-u)$. It is clear without further computation that the result of performing the inner four integrals must be a function of ρ alone (and certainly not of ϕ_s or ϕ_u, which do not appear in the integrands), say $h(\rho(s-u)) = K(s-u)$. We now assume that although Δt is a short interval (tending to zero) insofar as the slow phase error process is concerned, it is very long compared to the time scale of the noise processes N_1, N_2, X, or Y. It is therefore appropriate (and in keeping with the spirit of Viterbi's work, in which $K(s-u) = N_o \delta(s-u)$) to write

$$A_2 = K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} ds \int_{t-\Delta t/2}^{t+\Delta t/2} K(s-u)du$$

$$\approx K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} ds \int_{-\infty}^{\infty} K(\tau)d\tau$$

$$= K^2 \int_{-\infty}^{\infty} K(\tau)d\tau \triangleq K^2 B_o. \qquad (10)$$

Thus the Fokker-Planck equation (4) becomes $dp/d\phi = -2C/K^2B_o$; for symmetry, we take C=0, so that in the no-signal case the steady state mod-π density function of ϕ is uniform on $(-\pi/2,\pi/2)$.

5. THE STRONG-SIGNAL CASE, $A/\sigma \gg 1$

Once again denoting the numerator and denominator of the argument of the arctangent in eq.(1) by Y and X, we find them still to be independent Gaussian processes, each with covariance $\sigma^2\rho(\tau)$, but with mean values A sin ϕ and A cos ϕ respectively. Hence the analogue of eq.(8) for A_1 is

$$A_1 = -K \int_{-\pi/2}^{\pi/2} \theta d\theta \int_{-\infty}^{\infty} |r| e^{-\frac{r^2-2Ar\cos(\theta-\phi)+A^2}{2\sigma^2}} \frac{dr}{2\pi\sigma^2}$$

$$= -K \int_{-\pi/2}^{\pi/2} \frac{\theta d\theta}{2\pi\sigma^2} e^{-\frac{A^2\sin^2(\theta-\phi)}{2\sigma^2}}$$

$$\times\left\{\sqrt{2\pi}\,A\sigma\cos(\theta-\phi)\left[2\Phi\left(\frac{A\cos(\theta-\phi)}{\sigma}\right)-1\right]+2\sigma^2 e^{-\frac{A^2\cos^2(\theta-\phi)}{2\sigma^2}}\right\}$$

$$= -\frac{KA}{\sigma\sqrt{2\pi}} \int_{-\pi/2}^{\pi/2} \theta\cos(\theta-\phi)\left[2\Phi\left(\frac{A\cos(\theta-\phi)}{\sigma}\right)-1\right] e^{-\frac{A^2\sin^2(\theta-\phi)}{2\sigma^2}} d\theta, \tag{11}$$

where $\Phi(\cdot)$ is the unit normal distribution function. For $\frac{A}{\sigma}$ sufficiently large, the factor $(A/\sigma\sqrt{2\pi})e^{-\frac{A^2\sin^2(\theta-\phi)}{2\sigma^2}}$ behaves in the integral approximately as $\delta(\theta-\phi^*)$ where ϕ^* is as defined in section 2. Using the approximation $2\Phi(x)-1 \approx 1-2e^{-x^2/2}/x\sqrt{2\pi}$, valid for large positive x, eq.(11) can be integrated approximately to yield $A_1 \approx -K\phi^*$, a result fully in accord with expectations raised by eq.(3) for the noiseless case.

In computing A_2 according to eq.(6), the expecta-

tion, if taken inside the integrals over s and u, will clearly result in the sum of a "covariance" term and a "product of means" term; the contribution to A_2 of the product of means will evidently be

$$K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \int \phi^*(s)\phi^*(u)\,ds\,du = 0.$$

The integrations needed to compute the expectation will be at least as messy as that in eq.(11), so to simplify the anticipated necessary approximations we shall remove the mean values first. Hence we define

$$-\psi \triangleq \arctan \frac{(A+N_1)\sin\phi - N_2\cos\phi}{(A+N_1)\cos\phi + N_2\sin\phi} - \phi^*, \tag{12}$$

and note that $\psi = \arctan \frac{N_2}{A+N_1}$; the branch of the arctangent to be used is that between $-\pi/2+\phi^*$ and $\pi/2+\phi^*$ because of assumption A4. Eq.(6) therefore becomes

$$A_2 = K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \int ds\,du\; E\,\psi(s)\psi(u) \; . \tag{13}$$

Now we note that ψ does not depend on ϕ; so when computing the expectation as in eq.(9), $\phi(s)$ and $\phi(u)$ appear only in the limits of integration:

$$A_2 = K^2 \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \int ds\,du \int_{-\pi/2+\phi^*(s)}^{\pi/2+\phi^*(s)} \psi_s\,d\psi_s \int_{-\pi/2\phi^*(u)}^{\pi/2+\phi^*(u)} \psi_u\,d\psi_u \; Q(A,\rho,\sigma,\psi_s,\psi_u) \; , \tag{14}$$

where Q is the inner pair of integrals over r_s and r_u. When A/σ is very large, we expect that the phase error will be very small most of the time; therefore approximate expressions for A_2 can be found using a bivariate Maclaurin expansion of eq.(14) in $\phi^*(s)$ and $\phi^*(u)$. The zeroth order approximation, the only one we shall treat here, is of course found by evaluating eq.(14) at $\phi^*(s) = \phi^*(u) = 0$; following the reasoning leading to eq.(10), we then obtain

$$A_2 \triangleq K^2 C_o = K^2 \int_{-\infty}^{\infty} E\left[\psi(s)\psi(u) \mid \phi^*(s) = \phi^*(u) = 0\right] d(s-u). \quad (15)$$

Thus to this degree of approximation we have for the steady-state mod π density of ϕ

$$\phi^* p + \frac{C_o K}{2} \frac{dp}{d\phi} = 0,$$

the solution of which is a Gaussian density function with variance $C_o K/2$.

6. MEAN TIME TO LOSS OF LOCK

Since we have in the preceding sections treated only the steady-state distribution of the phase error modulo π, it is worthwhile further to calculate the mean time for the loop to lose lock; i.e., assuming ϕ to be initially zero, the expected time T for ϕ first to leave the interval $|\phi| < \pi/2$. Tausworthe (1967) has shown that $T(\phi_o)$ satisfies the ordinary differential equation

$$\frac{d^2T}{d\phi_o^2} + (2A_1/A_2) \frac{dT}{d\phi_o} + (2/A_2) = 0, \quad (16)$$

where ϕ_o is the initial value of ϕ and the boundary conditions are $T(\pm \pi/2) = 0$.

In the noiseless case ($A = 0$) we have $A_1 = 0$, $A_2 = K^2 B_o$; eq. (16) becomes

$$T'' = -2/K^2 B_o \quad (17)$$

whose solution is

$$T = (1/K^2 B_o)[(\pi/2)^2 - \phi_o^2]. \quad (18)$$

Thus the mean time to loss of lock is $T(0) = \pi^2/4B_o K^2$, an expression identical, in its dependence on the VCO gain K, to that found by Viterbi (1963) and Tausworthe (1967) for a conventional phase-locked loop.

When A/σ is very large, we have (approximately) $A_1 = -K\phi^*$, $A_2 = K^2 C_o$; since we are here concerned with ϕ in the interval $[-\pi/2, \pi/2]$, $\phi^* = \phi$ and eq.(16) becomes

$$T'' - (2\phi_o/KC_o)T' = -2K^2 C_o . \quad (19)$$

Since this equation is linear and of first order in T', a first integral is immediate.

$$T' = -2(\pi/K^3C_o)^{1/2} e^{\phi_o^2/KC_o}\left[\Phi\left(\frac{\phi_o}{\sqrt{KC_o/2}}\right) - \frac{1}{2}\right] . \qquad (20)$$

Expanding in infinite series (Abramowitz and Stegun (1964), formula 26.2.10), integrating term-by-term, and setting $\phi = 0$ yields the following rapidly convergent expression for the mean time to loss of lock:

$$T(0) = (2K^3C_o)^{-1/2} \sum_{o}^{\infty} (\pi/2)^{2n+2}\, n!/(2n+2)! \qquad (21)$$

T(0) for a conventional phase-locked loop operating at high signal-to-noise ratio decreases with K somewhat more rapidly than $K^{-3/2}$.

7. CONCLUSION

We have presented a preliminary and approximate analysis of a new quasi-linear phase-locked loop; much remains to be done however. Of particular interest will be the behavior of the loop at intermediate values of A/σ; we conjecture, for example, that since the noiseless results (section 2) show no dependence on A, the loop will display pronounced threshold effects as A/σ decreases. Only a more refined analysis than that we have given here is capable of answering this and other interesting questions.

REFERENCES

R. Tausworthe (1967) "Cycle Slipping in Phase-Locked Loops", I.E.E.E. Trans. on Communication Technology COM-15, No. 3.

V.I. Tikhonov (1959) "The Effects of Noise on Phase-Lock Oscillation Operation", Automatika i Telemekhanika 22, No. 9.

A.J. Viterbi (1963) "Phase-Locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques", Proc. I.E.E.E. 51, No. 12.

M. Abramowitz and I. Stegun (eds.) (1964) "Handbook of Mathematical Functions", NBS Applied Mathematics Series-55, Government Printing Office, Washington, D.C.

CONSTANT DIRECTIONS OF THE RICCATI EQUATION

DAVID RAPPAPORT AND RICHARD S. BUCY,
University of Southern California, Los Angeles

ABSTRACT

An examination is made of the discrete time autonomous matrix Riccati Equation

$$P_{n+1} = \phi' P_n \phi - \phi' P_n g(g' P_n g)^{\#} g' P_n \phi + H'H.$$

This equation occurs in a class of optimal control (filter) problems. The research problem is defined as a search for constant directions of the Riccati Equation. These are directions in which the Riccati Equation reaches an equilibrium point in finite time. The set of constant directions is shown to comprise a vector space. The dynamic dimension of the Riccati Equation is thereby reduced. Necessary and sufficient conditions for the existence of constant directions are determined. Applications to control and filtering are given.

INTRODUCTION

In the solution of optimal linear quadratic control problems and in optimal linear filtering problems with gaussian randomness, the matrix of quadratic form for performance index under optimal control, and the optimal error covariance matrix, satisfy nonlinear difference equations in discrete time, nonlinear differential equations in continuous time which are of the Riccati type. In references [5],[9]-[12], these equations are studied. For the particular class of optimal control problems in which the cost is not explicitly dependent on control, and the filtering problem in which the observations contain additive correlated noise (i.e., colored noise), it has been noted that it is possible to reduce the computational complexity of the optimal system; that is, elimination of instabilities, singularities, and reduction in the dimension of the optimal controller or filter.

Efforts to reduce the computational complexity of these dual control and filtering problems fall into two groups; the continuous time and the discrete time cases. The former case is found in [2],[3] and [4]. A common aspect of these approaches is the conversion by means of data differencing of a given system into a different, but equivalent form, which contains white observation noise, and for which the optimal filter is derivable. The discrete time case is considered in [6] and [8]. Both [6] and [8] reduce the size of the totally augmented state model by the number of observations.

The case studied here is the discrete time, autonomous single input(output)optimal control (filtering) problem. It is demonstrated that the relevant concept for reduction in computational complexity is that of constant directions of the Riccati Equation. In the control problem a consequence of this concept is necessary and sufficient conditions that the dead beat control [13]be optimal for a given initial system state. For the case of filtering, the optimal one-step predictor is shown to consist of two component parts; one,corresponding to the set of constant directions, being a delay line with constant tap gains, and the other utilizing the complete set of past outputs.

The results are obtained via consideration of the control problem. The results developed for the control case are then applied to optimal filtering by reason of duality. The results indicate that the discrete time data differencing techniques,[6], [8], are not necessary to accomplish reduction in computational complexity and in fact do not achieve the maximum possible.

This paper is based on a doctoral dissertation submitted to the University of Southern California by D. Rappaport. Research supported by AFOSR Grant AF-AFOSR-1244-67 and NASA Grant NGL-05-018-044, suppl. #3.

I. Statement of the Problem

A. The cost-free-control quadratic regulator problem (see [14], Chap. 13)

a) The motion of the autonomous system is specified by

$$\underline{x}_{n+1} = \phi \underline{x}_n + g\, q_n, \quad n=1, 2, \ldots, N \qquad \text{(II-1)}$$
$$\underline{x}_o = \underline{c}$$

where; $\underline{x}_n$ is column vector of dimension d; ϕ a nonsingular (d×d) matrix; g a (d×1) matrix; and q_n a scalar, the control variable.

b) The scalar on the interval [0,N] cost, not explicitly dependent on the control variable is

$$V(0,\underline{c};\Gamma,N) = \sum_{i=o}^{N-1} ||\underline{x}_i||^2_{H'H} + ||\underline{x}_N||^2_{\Gamma} \qquad \text{(II-2)}$$

where; H' is a (d×r) matrix $d \geq r$, and Γ is a (d×d) positive-semi-definite (≥ 0) matrix. The optimal cost is given as

$$V^o(0,\underline{c};\Gamma,N) = \min_{q_i, 0\leq i\leq N-1} \cdot V(0,\underline{c};\Gamma,N) = ||\underline{c}||^2_{P_N},$$

where

$$P_N = \phi' P_{N-1} \phi - \phi' P_{N-1} g (g' P_{N-1} g)^{\#} g' P_{N-1} \phi + H'H, \; P_o = \Gamma \qquad \text{(II-3)}$$

$\#$ denotes the penrose pseudo inverse (see [11]). Denote the solution to (II-3) by $\Pi(n,\Gamma)$.

Note that the autonomous nature of the system implies that the cost is the same on any time interval of the same duration as [0,N]. This enables the fixing of the initial time, here conveniently taken as zero, while the final time N is moved ahead as optimization on intervals of increasing length are considered.

B. The correlated noise problem (see [4]):

a) The motion of the autonomous system is specified by

$$\underline{x}_{n+1} = \psi \underline{x}_n + G\, \underline{u}_n \qquad \text{(II-4)}$$
$$\underline{x}_o = \underline{c}$$

with scalar output

$$z_n = h\, x_n, \qquad \text{(II-5)}$$

where; $\underline{x}_n$ is a (d×1) matrix, ψ a (d×d) nonsingular matrix, G a (d×r) matrix, $d \geq r$, h a (1×d) matrix, $\underline{u}_n$ is a vector zero-mean gaussian white noise sequence, independent of x_n with covariance $E\, \underline{u}_i\, \underline{u}_j' = \delta_{ij} I_r$, $\underline{c}$ is a zero-mean gaussian random variable independent of $\underline{u}_n$, with covariance $E\, \underline{c}\, \underline{c}' = \Gamma$.

(II-4) and (II-5) are the dynamic equations for the so-called totally augmented system.

b) As is well-known, ([9], [11]), the minimum variance, unbiased one-step predictor estimate is given by the conditional expectation

$$\hat{\underline{x}}_{n+1/n} = E(x_{n+1}/z_i \;\; 0 \leq i \leq n)$$

where

$$\hat{\underline{x}}_{n+1/n} = \phi\, \hat{\underline{x}}_{n/n-1} + \phi\, P_n\, h' (h\, P_n\, h')^{\#} (z_n - h\, \hat{\underline{x}}_{n/n-1}) \text{ a.s.} \qquad \text{(II-6)}$$

# denotes the penrose-pseudo inverse and the error covariance function $P_n = E(\underline{x}_n - \hat{\underline{x}}_{n/n-1})(\underline{x}_n - \hat{\underline{x}}_{n/n-1})'$ satisfies the nonlinear Riccati difference equation

$$P_{n+1} = \Psi P_n \Psi' - \Psi P_n h'(h P_n h')^{\#} h P_n \Psi' + GG', \; P_o = \Gamma. \qquad \text{(II-7)}$$

C. The optimal control and optimal filtering problems just outlined are duals of one another under the mappings

$$h \longleftrightarrow g', \quad G \longleftrightarrow H', \quad \phi \longleftrightarrow \Psi' . \qquad \text{(II-8)}$$

An immediate result under this duality transformation is that the Riccati equations (II-3) and (II-7) map into each other. See references [5], [9] and [11] for more information on duality.

D. An important concern in the implementation of the optimal filter and optimal controller is the dynamic dimension of the Riccati Equation, i.e., the number of independent varying directions, and the elimination of constant singularities. This paper is restricted to the case of the scalar input (control) or scalar output (filtering), although the sufficiency conditions are in actuality not so constrained. The problem to be confronted can be phrased as follows:

> "To find the set of all directions C for which $||\underline{c}||^2_{\Pi(n,\Gamma)}$, $\underline{c} \varepsilon C$, reaches an equilibrium point in finite time, for all initial conditions Γ, where $\Pi(n,\Gamma)$ is the solution of equation (II-3)."

II. Preliminary Results:

A. DEFINITION: $\underline{c}$ is a constant direction of equation (II-3) if there exists some finite N such that for $\forall\, n \geq N$ and $\forall\, \Gamma_1 \geq 0$, $\Gamma_2 \geq 0$, $||\underline{c}||^2_{\Pi(n,\Gamma_1)} = ||\underline{c}||^2_{\Pi(n,\Gamma_2)}$. Let the set of all constant directions be denoted by C.

The requirement that $\underline{c}$ be a constant direction for all $\Gamma \geq 0$ implies that the constant directions are independent of Γ. The next theorem gives a more physical interpretation of a constant direction.

B. THEOREM: $\underline{c} \varepsilon C$ of (II-3) for $\forall\, n \geq N$ iff $\underline{c}$ can be taken optimally to $\underline{o}$ in the interval $[0,N]$ for some Γ.

PROOF: Let $\underline{c}, \underline{x}_1, \ldots, \underline{x}_{N-1}, \underline{x}_N = \underline{0}$ be an optimal path which takes $\underline{c}$ to $\underline{o}$ on $[0,N]$ for some Γ_1. Therefore from (II-2)

$$V^o(0,\underline{c};\Gamma_1,N) = \sum_{i=o}^{N-1} ||\underline{x}_i||^2_{H'H} \quad \text{where } \underline{x}_o = c.$$

Let $\underline{x}^*_o = \underline{c}, \underline{x}^*_1, \ldots, \underline{x}^*_{N+J}$, $J \geq 0$ be an optimal path on $[0,N+J]$ for Γ_2 arbitrary. Hence

$$V^o(0,\underline{c};\Gamma_2,N+J) = \sum_{i=o}^{N+J-1} ||\underline{x}^*_i||^2_{H'H} + ||\underline{x}^*_{N+J}||^2_{\Gamma_2}$$

Let controls $q_i = 0$, $i = N, \ldots, N+J-1$. Hence $\underline{c}, \underline{x}_1, \ldots, \underline{x}_{N+J}$, where $\underline{x}_i = 0$, $i \geq N$ is a continuation of the trajectory $\underline{c}, \underline{x}_1, \ldots, \underline{x}_N$. By optimality on $[0, N+J]$

$$\sum_{i=o}^{N+J-1} ||\underline{x}^*_i||^2_{H'H} + ||\underline{x}^*_{N+J}||^2_{\Gamma_2} \leq \sum_{i=o}^{N+J} ||\underline{x}_i||^2_{H'H} = \sum_{i=o}^{N-1} ||\underline{x}_i||^2_{H'H}, \; J \geq 0 .$$

Therefore $\sum_{i=o}^{N-1} ||\underline{x}_i^*||^2_{H'H} \leq \sum_{i=o}^{N-1} ||\underline{x}_i||^2_{H'H}$.

But optimality on [0,N] implies

$$\sum_{i=o}^{N-1} ||\underline{x}_i||^2_{H'H} = \sum_{i=o}^{N+J} ||\underline{x}_i||^2_{H'H} \leq \sum_{i=o}^{N-1} ||\underline{x}_i^*||^2_{H'H} \leq \sum_{i=o}^{N+J-1} ||\underline{x}_i^*||^2_{H'H} + ||\underline{x}_{N+J}^*||^2_{\Gamma_2}, J \geq 0.$$

Hence $V^o(0,\underline{c};\Gamma_1,N) = V^o(0,\underline{c};\Gamma_2,N+J)$, $J \geq 0$, Γ_2 arbitrary implies that $\underline{c} \varepsilon C$.

Conversely, suppose that $\underline{c} \varepsilon C$ of (II-3) for $\forall\, n \geq N$. Then by Definition A $V^o(0,\underline{c};\Gamma_1,n) = V^o(0,\underline{c};\Gamma_2,n), \Gamma_1 \neq \Gamma_2$ for $\forall\, n \geq N$.

Let $\Gamma_1 = 0$, $\Gamma_2 > 0$, let $\underline{x}_i^{(1)}$ and $\underline{x}_i^{(2)}$ denote respectively the optimal paths on [0,N] emanating from $\underline{x}_o^{(1)} = \underline{x}_o^{(2)} = \underline{c}$ for Γ_1 and Γ_2 .

Then $\sum_{i=o}^{N-1} ||\underline{x}_i^{(1)}||^2_{H'H} = \sum_{i=o}^{N-1} ||\underline{x}_i^{(2)}||^2_{H'H} + ||\underline{x}_N^{(2)}||^2_{\Gamma_2}$ and assuming $\underline{x}_N^{(2)} \neq \underline{0}$ implies that $\sum_{i=o}^{N-1} ||\underline{x}_i^{(1)}||^2_{H'H} > \sum_{i=o}^{N-1} ||\underline{x}_i^{(2)}||^2_{H'H}$.

But $\underline{x}_i^{(1)}$, i = 0, ..., N-1 is an optimal path on [0,N-1] since

$$V^o(0,\underline{c};\, 0,N) = V^o(0,\underline{c};\, H'H,\, N-1) = \sum_{i=o}^{N-1} ||\underline{x}_i^{(1)}||^2_{H'H} .$$

However

$$V^o(0,\underline{c};\, H'H,\, N-1) > \sum_{i=o}^{N-1} ||\underline{x}_i^{(2)}||^2_{H'H} \text{ , a contradiction !}$$

Hence $\underline{x}_N^{(2)} = \underline{0}$ implies that $\underline{x}_n$ can be taken optimally to $\underline{0}$ for some Γ on [0,N]. ■

Remark: By the autonomous nature of the system if $\underline{c}$ can be taken optimally to $\underline{0}$ on [0,N], then it can also be driven optimally to the origin on [0,N] all $n \geq N$.

C. DEFINITION OF MOMENT: The i^{th} moment of system (II-1) is defined to be $H\, \phi^i\, g = u_i$, i an integer.

Remark: Moments are coordinate system invariant properties of a system.

D. LEMMA: $\phi^{-1} g$ is a constant direction of (II-3) for $\forall\, n \geq 1$, and $\Pi(n,\Gamma)\, \phi^{-1} g = H'H\, \phi^{-1} g = H' u_{-1}$ for $\forall\, \Gamma$.

PROOF: Follows immediately from (II-3). ■ Lemma D means that (II-3) always has at least one constant direction. Let $R^i = \text{span}\{\phi^{-1}g,..,\phi^{-i}g\}$.

Let $m \leq d$ be the largest positive integer such that $\{\phi^{-1}g,\ldots\, \phi^{-m}g\}$ are linearly independent.

Note that R^m is the space of completely controllable states (see [13]). It follows that Lemma E and Corollary F can be easily verified.

E. LEMMA: The vector $\phi^{-i}\, g \varepsilon R^m$ for all integers i .

F. COROLLARY: Every consecutive "string" of vectors $\{\phi^i g,..,\phi^{i+(m-1)} g\}$, all i are linearly independent.

G. LEMMA: Let $\underline{c} \varepsilon C$ of equation (II-3) for $\forall\ n \geq N$, where N is the first $n \ni$

$$||\underline{c}||^2_{\Pi(n,\Gamma_1)} = ||\underline{c}||^2_{\Pi(n,\Gamma_2)}$$

for $\forall\ n \geq N$, $\forall \Gamma_1 \geq 0$, $\Gamma_2 \geq 0$. Let $\underline{x}_i$, $i = 0,\ldots, N$, $\underline{x}_o = \underline{c}$ be an optimal path on $[0,N]$. Then $\underline{x}_i$, $i = 0,\ldots, N-1 \varepsilon C$ of (II-3).

PROOF: By Theorem B $\underline{x}_N = 0$. By the principle of optimality [1] the sequences $\underline{x}_i, \underline{x}_{i+1},\ldots, \underline{x}_N = 0$ $i = 0, 1,\ldots,N$ are optimal trajectories on the respective intervals $[i,N]$. Hence by Theorem B these $\underline{x}_i \varepsilon C$ of (II-3)

H. LEMMA: C, the set of constant directions is a subset of R^m.

PROOF: In (II-1) only completely controllable states (see [13]) can be taken to $\underline{0}$. Hence by Theorem B, $C \varepsilon R^m$. ■

Let $\underline{c} \varepsilon C$ be a state which can be carried optimally to $\underline{0}$ in exactly N steps. Then by Lemma H

$$\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, a_i \quad \text{where } k \leq m \text{ and } a_k \neq 0. \tag{III-1}$$

Then from equation (II-1) it follows that

$$\left.\begin{aligned} \underline{x}_1 &= \phi\, \underline{c} + g\, q_o \\ &\vdots \\ \underline{x}_{N-1} &= \phi^{N-1} \underline{c} + \phi^{N-2} g\, q_o + \ldots + g\, q_{N-2} \neq 0 \\ \underline{x}_N &= \underline{0} \end{aligned}\right\} \tag{III-2}$$

I. LEMMA: $\underline{c} \varepsilon C$ cannot be taken to $\underline{0}$ in less than k steps, i.e., in a proper subinterval of $[0,k]$.

PROOF: Follows from (III-2) and Corollary F.

J. LEMMA: Let $\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, a_i$, $k \leq m$, $a_k \neq 0$, $\underline{c} \varepsilon C$. Then $\underline{x}_{N-i}$ (equation (III-2)) εC, $i = 1,\ldots, k$ and are linearly independent.

PROOF: By Lemma G, $\underline{x}_i$, $i=o,\ldots, N-1$ in (III-2) are constant directions of (II-3).

$\underline{x}_N = 0 = \phi\, \underline{x}_{N-1} + g\, q_{N-1}$, $q_{N-1} \neq 0$ since $\underline{x}_{N-1} \neq 0$ by hypothesis. An immediate consequence of equation (II-1) is that

$$\underline{x}_{N-i} = -\phi^{-i} g\, q_{N-1} - \ldots - \phi^{-1} g\, q_{N-i},\ i=1,\ldots, N, \tag{III-3}$$

and by Lemma G, $\underline{x}_{N-i}$ is a constant direction for $n \geq i$.

Note that since $q_{N-1} \neq 0$, $\underline{x}_{N-i}$, $i=1,\ldots, k$, where k is defined by equation (III-1), are k linearly independent vectors belonging to R^m; namely, $\underline{c}$ in equation (III-1) belonging to C generates k linearly

independent constant directions. ■

For notational convenience in equation (III-3) let

$$\underline{x}_{N-i} = \underline{y}_i \qquad i=1,\ldots,k. \tag{III-4}$$

K. LEMMA: If $\underline{y}_i$ (in III-4) $i = 1,\ldots, k \leq m \leq d$ are linearly independent constant directions for all $n \geq i$, where $\underline{y}_i = \phi^{-i}g\, a_1 + \ldots + \phi^{-1}g\, a_i, a_1 \neq 0$, then

$$W^{i-1}\begin{bmatrix} \sum_{j=o}^{i-2} u_j'\, u_{-(i-j-1)} \\ \vdots \\ u_o'\, u_{-2} + u_1'\, u_{-1} \\ u_o'\, u_{-1} \end{bmatrix} = \underline{0} \text{ for } 2 \leq i \leq k, \text{ where } u_i'\, u_j = g'\phi^{-i'}H'H\,\phi^j g$$

$$\Longrightarrow g'\Pi(n-1,\Gamma)\phi^{-i}g = 0 \quad n \geq i+1,\ i=1,..,k-1, \text{ all } \Gamma$$

$$\Longrightarrow \Pi(n,\Gamma)\phi^{-i}g = \sum_{j=1}^{i} \phi^{j-1'}H'H\,\phi^{-i+j-1}g \sum_{j=1}^{i} \phi^{j-1'}H'\, u_{-(i-j+1)}$$

constant for all $n \geq i$, $i = 1,\ldots, k$, all Γ .

Remark: Notice that $u_i'\, u_j$ is just the inner product of the i^{th} and j^{th} moments (see Definition C) and that $Wi = 0 \Rightarrow W^{i-1} = \underline{0}$.

PROOF: Case of $i = 2$.

Let $||y_2||^2_{\Pi(n,\Gamma)}$ be constant for $n \geq 2$, all Γ . From (II-3) $\Pi(n,\Gamma)\, \underline{y}_2 = \Pi(n,\Gamma)(\phi^{-2}g\, a_1 + \phi^{-1}g\, a_2)$. By Lemma C

$$\Pi(n,\Gamma)\, \underline{y}_2 = H'H\,\phi^{-2}g\, a_1 + \phi'H'H\,\phi^{-1}g\, a_1 + H'H\,\phi^{-1}g\, a_2 - \phi'\Pi(n-1,\Gamma)g(g'\Pi(n-1,\Gamma)g)^{\#}\, u_o'\, u_{-1}\, a_1.$$

Therefore for $n \geq 2$

$$||\underline{y}_2||^2_{\Pi(n,\Gamma)} = \text{constant(independent of } \Gamma) - a_1^{\,2}(u_o'\, u_{-1})^2 (g'\Pi(n-1,\Gamma)\, g)^{\#}$$

Suppose $u_o'\, u_{-1} \neq 0$, then

$$0 = ||\underline{y}_2||^2_{\Pi(n+1,\Gamma_1)} - \Pi(n,\Gamma_2)$$

$$= a_1^{\,2}\,(u_o'\, u_{-1})^2\,[(g'\Pi(n-1,\Gamma_2)\, g)^{\#} - (g'\Pi(n,\Gamma_1)\, g)^{\#}],\ n \geq 2$$

for all Γ_1, Γ_2. Therefore, since $a_1 \neq 0$, and $(g'\Pi(n,\Gamma)\, g)^{\#}$ is a unique scalar quantity $g'\Pi(n-1,\Gamma_2)g = g'\Pi(n,\Gamma_1)g$ for $n \geq 2$, all Γ_1, Γ_2.
From Lemma E and Corollary F $g = \alpha_m\, \phi^{-m}g + \ldots + \alpha_1\, \phi^{-1}g, \alpha_m \neq 0$.
By Lemma I and Theorem B g cannot be a constant direction for $n < m$.

Since $2 \leq m \leq d$, then $g'\Pi(1,\cdot)\,g = g'\Pi(2,\cdot)\,g = \ldots$ is a contradiction ! Hence $u_o' u_{-1} = 0$, and by Lemma C $u_o' u_{-1} = g'\Pi(n-1,\cdot)\phi^{-1} g = 0$ $n \geq 2$. Also equation (II-3) gives $\Pi(n,\cdot)\phi^{-2} g = \phi' H'H \phi^{-1} g + H'H \phi^{-2} g$ $n \geq 2$. Therefore the lemma is valid for $i = 2$.

To employ mathematical induction: assume the lemma holds for the case of $j < k$, i.e., y_i, $i = 1,\ldots, k$ linearly independent constant directions imply $W^{j-1} = \underline{0}$, etc. Thus, $\Pi(n,\Gamma)y_{j+1} = \Pi(n,\Gamma)(\phi^{-(j+1)} g\, a_1 + \ldots + \phi^{-1} g\, a_{j+1})$ by the induction hypothesis becomes

$$\Pi(n,\Gamma)y_{j+1} = \Pi(n,\Gamma)\phi^{-(j+1)} g\, a_1 + \text{constant (independent of } \Gamma\text{), for } n \geq j.$$

Substituting equation (II-3) gives

$$\Pi(n,\Gamma)y_{j+1} = \text{constant} + [\phi'\Pi(n-1,\Gamma)\phi^{-j} g - \phi'\Pi(n-1,\Gamma)g(g'\Pi(n-1,\Gamma)g)^{\#} g'\Pi(n-1,\Gamma)\phi^{-j} g]\, a_1$$

The induction hypothesis implies that $\Pi(n,\Gamma)y_{j+1} = \text{constant}$

$$-\phi'\Pi(n-1,\Gamma)g(g'\Pi(n-1,\Gamma)g)^{\#}\left(\sum_{k=o}^{j-1} u'_k u_{-(j-k)}\right) a_1, \text{ for } n \geq j+1$$

Let $\sigma_j = \sum_{k=o}^{j-1} u'_k u_{-(j-k)}$. Further manipulation results in

$$||y_{j+1}||^2_{\Pi(n,\Gamma)} = \text{constant(independent of } \Gamma\text{)} - a_1^2\, (\sigma_j)^2 (g'\Pi(n-1,\Gamma)g)^{\#}, n \geq j+1.$$

Therefore for all Γ_1, Γ_2

$$0 = ||y_{j+1}||^2_{\Pi(n+1,\Gamma_1)} - \Pi(n,\Gamma_2)$$
$$= a_1^2(\sigma_j)^2[(g'\Pi(n-1,\Gamma_2)g)^{\#} - (g'\Pi(n,\Gamma_1)g)^{\#}], \text{ for } n \geq j+1.$$

Suppose $\sigma_j \neq 0$. Since $a_1 \neq 0$, and $(g'\Pi(n,\cdot)g)^{\#}$ is a unique scalar quantity $g'\Pi(n-1,\Gamma_2)g = g'\Pi(n,\Gamma_1)g$ for $n \geq j+1$, all Γ_1, Γ_2; i.e., $g'\Pi(j,\cdot)g = g'\Pi(j+1,\cdot)g = \ldots$ As in the case of $j = 2$, g cannot be a constant direction for $n < m$. Since $j < k \leq = m$, $g'\Pi(j,\cdot)g = g'\Pi(j+1,\cdot)g = \ldots$, is a contradiction ! Hence $\sigma_j = 0$, i.e., $W^j = \underline{0}$.

But this in turn implies that

$$g'\Pi(n-1,\cdot)\phi^{-j} g = \sum_{k=o}^{j-1} u'_k u_{-(j-k)} = 0, \text{ for } n \geq j+1,$$

further implying

$$\Pi(n,\Gamma)\,\phi^{-(j+1)} g = \sum_{k=1}^{j+1} \phi^{k-1\prime} H' u_{-(j+1)+k-1} \quad \text{for } n \geq j+1, \text{ all } \Gamma. \blacksquare$$

L. LEMMA: $\Pi(n,\Gamma)\phi^{-i} g$ constant independent of Γ for all $n \geq i$, $i = 1\ldots$, k implies that $\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, a_i$, a_i arbitrary, is a constant direction of (II-3).

PROOF: $\Pi(n,\Gamma)\,\underline{c}$ = constant for $n \geq k$, Γ independent, implies

$$||\underline{c}||^2_{\Pi(n,\Gamma)} = ||\underline{c}||^2_{\Pi(n,\Gamma_1)} = ||\underline{c}||^2_{\Pi(n,\Gamma_2)}, \text{ all } n \geq k, \text{ all } \Gamma_1, \Gamma_2.$$

Hence $\underline{c}$ is a constant direction by Definition A. ■

MAIN RESULTS

Summarizing the results of Lemmas J, K and L gives

M. THEOREM: Let $c = \sum_{i=1}^{k} \phi^{-i} g\, a_i$, $a_k \neq 0$, $k \leq m$, then $\underline{c} \in C$ the set of constant directions $\Leftrightarrow W^{k-1} = \underline{0} \Leftrightarrow g'\Pi(n-1,\Gamma)\phi^{-i}g = 0$, $n \geq i+1$, $i=1,\ldots,k-1$, all $\Gamma \Leftrightarrow \Pi(n,\Gamma)\phi^{-i}g = \sum_{j=1}^{i} \phi' {}^{j-1} H'H\, \phi^{-(i+1)+j} g$ for $n \geq i, i=1,\ldots,k$, all Γ. ■

The following theorem is the principal result. It provides a complete characterization of C and a method for the quick determination of all constant directions.

N. THEOREM: $W^{k-1} = \underline{0}$, $W^k \neq \underline{0} \Leftrightarrow C = R^k$.

PROOF: Assume $W^{k-1} = \underline{0}$, $W^k \neq 0$. By Theorem M, $W^{k-1} = \underline{0}$ implies that $\Pi(n,\Gamma)\phi^{-i}g$ = constant all $n \geq i$, $i=1,\ldots,k$, all Γ. Therefore $\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, \gamma_i$, γ_i arbitrary, is a constant direction, i.e., $R^k \subset C$. By Lemma H, $C \subset R^m$. Let $\underline{\alpha} \in R^j$, $k \leq j \leq m$, $\underline{\alpha} \notin R^k$, thereby implying that $\underline{\alpha} = \sum_{i=1}^{\ell} \phi^{-i} g\, \gamma_i$, $k < \ell \leq j$, $\gamma_\ell \neq 0$. Assume $\underline{\alpha} \in C$. By M this means that $\underline{0} = W^{\ell-1}$, which further implies $W^k = \underline{0}$. A contradiction ! Hence $\underline{\alpha}$ is not a constant direction, proving $C = R^k$. Conversely, let $\underline{c} = \phi^{-k} g$. By M, $W^{k-1} = \underline{0}$. Let $\underline{\alpha} = \phi^{-(k+1)} g$. By M, $\underline{\alpha}$, not a constant direction implies $W^k \neq \underline{0}$. ■

Remark: The set of all constant directions is a vector space.

O. COROLLARY: If $\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, a_i, a_k \neq 0$ is a constant direction, $\Pi(n,\Gamma)\underline{c}$ is a constant independent of Γ for all $n \geq k$, and not constant for $0 \leq n < k$, for all Γ.

PROOF: By Theorem B and Lemma I. ■

Remark: From O it is obvious that the Riccati Equation (II-3) reaches equilibrium points in the constant directions in m or less steps, where m is the dimension of the space of controllable states.

Theorem M indicates that an apparently stronger definition of a constant direction of the Riccati Equation (II-3) is in actuality equivalent to the definition given in A; that is, the definition "$\underline{c}$ is a constant direction of (II-3) for all $n \geq N$ if the vector $\Pi(n,\Gamma)\,\underline{c}$ is a constant independent of Γ, for all $n \geq N$," is equivalent to A.

Implications of the Main Results: The following theorem describes when dead-beat control optimal (see [13]).

P. THEOREM: $\underline{c} = \sum_{i=1}^{k} \phi^{-i} g\, a_i$, $a_k \neq 0$, $k \leq m$ is a constant direction if and only if $\underline{c}$ can be taken optimally to $\underline{0}$ in exactly k steps by the dead-beat control, $q_i = a_{i+1}$, $i=0,\ldots,k-1$, for some Γ.

PROOF: If; follows immediately from Theorem B. Conversely, by

Lemma I $\underline{x}_i \neq \underline{0}$, i=0,..., k-1, where $x_o = \underline{c}$. By Corollary O and Theorem B, $\underline{c}$ is a constant direction only if $\underline{x}_k = \underline{0}$. By Corollary G the vectors $g,\ldots, \phi^{k-1}g$ are linearly independent. Since

$$\underline{x}_k = g(a_k + q_{k-1}) + \phi\, g(a_{k-1} + q_{k-2}) + .. + \phi^{k-1} g(a_1 + q_o)$$

$\underline{x}_k = 0$ only if $q_i = a_{i+1}$ i=o, .., k-1 . ■

Once the Riccati Equation (II-3) reaches the equilibrium point in a constant direction, any further calculation of the value of P_n in that direction would appear redundant. This motivates the following discussion.

Q. DEFINITION: Let ℓ be the largest positive integer such that $\phi^{-1}g$, $\ldots,\phi^{-\ell}g$ are constant directions of equation (II-3). Then dynamic rank of the Riccati Equation (II-3) = d-ℓ, where d = system dimension.

Consider that the system (II-1) is in the coordinate system with basis vectors $[\phi^{-1}g,\ldots,\phi^{-\ell}g, L]$,and d-$\ell$ is the dynamic rank of the corresponding Riccati Equation. It should be clear from Corollary O that after a transitory period of length ℓ, $\Pi(n,\Gamma)$ can be partitioned as

$$\Pi(n,\Gamma) = \begin{bmatrix} A & B \\ B' & W(n,\Gamma) \end{bmatrix} \quad , n \geq \ell \text{ where} \qquad (V\text{-}1)$$

$W(n,\Gamma)$ is a (d-ℓ) x (d -ℓ) matrix, and A and B are constant matrices independent of Γ. From equation (II-3) it follows that $W(n,\Gamma)$ can be computed recursively in terms of itself, and the constant matrices A and B by means of a (d-ℓ)×(d-ℓ) matrix Riccati type equation. Note that all Γ-independent constant singularities are eliminated from the iteration by this reduction since they are a subset of the set of constant directions.

Applications to Optimal Linear Filtering:

Under Mapping II-8, Theorems M and N can be dualized for the system (II-4) - (II-7). It will now be demonstrated that the dual of an optimal trajectory starting at $\underline{c}$ and arriving $\underline{0}$ in finite time period [0,k], is the filter direction $\underline{c}$ for which the best one-step predictor estimate is a constant linear combination of $z_{n-(k-1)}$, for $n \geq k-1$, $1 \leq k \leq d$. Note that in the proof below the estimate equations hold almost surely.

R. THEOREM: $\underline{\lambda}_k = \sum_{i=1}^{k} h\psi^{-i}\, a_i$ (a row vector), $a_k \neq 0$ is a constant direction of the Riccati Equation (II-7) if and only if

$$\underline{\lambda}_k \cdot \hat{\underline{x}}_{n+1/n} = \sum_{i=1}^{k} z_i\, a_{n-i+1} \text{ a.s., all } n \geq k .$$

(· denotes the vector inner product.)

PROOF: Assuming that $\underline{\lambda}_k$ is a constant direction of (II-7), by the dual of Theorem M and (II-6)

$$\underline{\lambda}_k \cdot \hat{\underline{x}}_{n+1/n} = \underline{\lambda}_k \cdot \psi \hat{\underline{x}}_{n/n-1} + a_1 h\Pi(n,\Gamma)h' (h\Pi(n,\Gamma)h')^{\#} (z_n - h\hat{\underline{x}}_{n/n-1}), \text{ for } n \geq k .$$

Assume $h\Pi(n,\Gamma)h' > 0$, then $\underline{\lambda}_k \cdot \hat{\underline{x}}_{n+1/n} = a_1\, z_n + \underline{\lambda}_{k-1} \cdot \hat{\underline{x}}_{n/n-1}$, where $\underline{\lambda}_{k-1} = \sum_{i=1}^{k-1} a_{i+1} h\psi^{-i}$, This is also true for $h\Pi(n,\Gamma)h' = 0$.

By the duals of Theorems M and N $\underline{\lambda}_{k-1}$ is also a constant direction of (II-7), and it can be shown that the desired implication follows inductively. Conversely, assume that

$$\underline{\lambda}_{k+1} \cdot \hat{\underline{x}}_{n+1/n} = \sum_{i=1}^{k} a_i\, z_{n-i+1}, \quad \text{all } n \geq k.$$

Note that from equations (II-4) and (II-5) it can be shown inductively that

$$h\,\psi^{-i}\, x_{n+1} = z_{n-i+1} + h\,\psi^{-1}\, G\, u_{n-i+1} + \ldots + h\,\psi^{-i}\, G\, u_n .$$

Hence $E\left[\sum_{i=2}^{k} a_i(h\psi^{-1}G\,\underline{u}_{n-i+1}+\ldots+h\psi^{-(i-1)}G\underline{u}_{n-1})\,|\,z_j,\ j \le n\right]=0$ all $n \ge k$.

By the smoothing property of conditional expectation and letting $j = i-1$, $n = m+1$,

$$E\left[\sum_{j=1}^{k-1} a_{j+1}(h\psi^{-1}G\underline{u}_{m+1-j}+\ldots+h\psi^{-j}G\underline{u}_m)\,|\,z_i, i \le m\right] = 0, \text{ all } m \ge k-1.$$

Hence

$$\underline{\lambda}_{k-1}\cdot\hat{\underline{x}}_{n+1/n} = \sum_{i=1}^{k-1} a_{i+1}\,z_{n-i+1}, \text{ all } n \ge k-1.$$

Inductively it follows that

$\underline{\lambda}_j\cdot\hat{\underline{x}}_{n+1/n} = \sum_{i=1}^{k-j} a_{i+1}\,z_{n-j+1}$, all $n \ge j$, where $\underline{\lambda}_j = \sum_{i=1}^{k-j} h\phi^{-i}\,a_{i+j}$.

The $\underline{\lambda}_j$, $j = 1,\ldots,k$, span the vector space R^k = span of $\{h\psi^{-1},..,h\phi^{-k}\}$ and for any $\underline{\lambda}\in R^k$, i.e., $\underline{\lambda} = \sum_{i=1}^{k} h\psi^{-i}\,b_i$, $\underline{\lambda}\cdot\hat{\underline{x}}_{n+1/n} = \sum_{i=1}^{k} b_i\,z_{n-i+1}$, $n \ge k$.

Since $h\psi^{-i}\in R^k$, $1 \le i \le k$, $h\psi^{-i}\,\hat{\underline{x}}_{n+1/n} = h\psi^{-(i-1)}\hat{\underline{x}}_{n/n-1} - h\psi^{-(i-1)}\Pi(n,\Gamma)$

$$h'(h\Pi(n,\Gamma)h')^{\#}(z_n - h\hat{\underline{x}}_{n/n-1}).$$

Therefore, for $2 \le i \le k$, $n \ge i-1$

$$z_{n-i+1} = z_{n-i+1} - h\psi^{-(i-1)}\Pi(n,\Gamma)h'(h\Pi(n,\Gamma)h')^{\#}(z_n - h\hat{\underline{x}}_{n/n+1}) .$$

Therefore, $h\psi^{-(i-1)}\Pi(n,\Gamma)h' = 0$, $2 \le i \le k$, $n \ge i-1$, all Γ. Hence by dual of Theorem M $\underline{\lambda}_k$ is a constant direction.

In sum, Theorem R states that one-step prediction in constant directions merely corresponds to a finite length delay line with constant tap gains. One-step prediction in nonconstant directions has the usual dynamic feedback form with varying gains in general.

REFERENCES

1. Bellman, R.E., "Dynamic Programming, Princeton Univ.Press, ('57).
2. Bryson,A.E.,Jr.,and Johansen,D.E.,"Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise,"IEEE Trans.,Vol. AC-10,pp 4-10, 1/65.
3. Bucy,R.S.,"Optimum Finite Time Filters for a Special Non-Stationary Class of Inputs,"Internal Rep.B.B.D.600,3/31/59,John Hopkins Appl.Phy.Lab.
4. Bucy,R.S.,"Optimal Filtering with Correlated Noise,"J.Math.&Appl.20('67).
5. Bucy,R.S.,& Joseph,P.D.,"Filtering for Stochastic Processes with Applications to Guidance,"Interscience, New York ('68).
6. Deyst,J.J., Masters Thesis, Aerospace Dept. M.I.T., '64.
7. Henrikson,L.J., Ph.D. Thesis,"Sequentially Correlated Measurement Noise with Applications to Inertial Navigation", Havard Univ., '67.
8. Kalman, R.E.," A New Approach to Linear Filtering and Prediction Problems," J. Basic Eng.,Trans. ASME, 82D, pp 34-45, 3/60.
9. Kalman, R.E., and Bucy R.S.,"New Results in Linear Filtering and Prediction Theory," J. Basic Eng., Trans.ASME, 83D, pp 95-108, 3/61.
10. Kalman,R.E.,"New Methods in Wiener Filtering Theory,"Proc.First Symposium on Engineering Applications of Random Function Theory, Purdue Univ.,11/60, pp 270-385, Wiley, New York ('63).
11. Kalman,R.E., "Contributions to the Theory of Optimal Control,"Bol.Soc. Math Mexicana, pp 102-119, '60.
12. Kalman, R.E.,"On the General Theory of Control Systems,"Proc. of the First Int'l Congress of the Int'l Fed.of Autom.Cont.,Vol.I, Moscow,'60.
13. Kalman,R.E., Englar,T.S,,Bucy,R.S.,"Fundamental Study of Adaptive Control Systems," Techn. Documentary Rep. No.ASD-TR-61-27,Vol.II, 10/64.

ON THE CONTROLLABILITY OF DISCRETE-TIME-VARYING LINEAR SYSTEMS

ANTHONY J. GRAMMATICOS
Department of Electrical Engineering
New York University
Bronx, New York 10453

ABSTRACT

A simplified sufficient condition for controllability of discrete-time-varying linear systems is derived. The analogy between this condition and a known similar condition for the continuous-time-varying case is pointed out. An illustrative example is given.

I. INTRODUCTION

The essence of the concept of controllability is that any desired changes of the state can be effected in finite time by using some appropriate input.

Controllability is an important characterizing property of dynamical systems and plays an important role in optimal control and network theory.

A formal definition of controllability was given in [1,2]*. This was followed by two self-contained papers on the subject, one of which examines the continuous-time case [3] and the other the discrete-time case [4]. The concept of output controllability was introduced in the last two papers.

General conditions for controllability of linear systems are given in the above papers. These conditions depend on the system fundamental matrix which in general, is not available. It was only in the time-invariant case where it was possible to express the controllability condition in terms of the coefficient matrices of the system and hence simplify them.

A simplified sufficient condition for controllability of continuous-time-varying linear systems was given in [5]. The condition was expressed in terms of the system matrices and a finite number of their derivatives by means of a recursive dfiferential relation. This controllability condition has found applications in the synthesis of time-variable unilateral networks, on the one hand [5]. On the other hand, the recursive relation on which the condition is based, appears to play a central role in the theory of canonical decomposition of time-varying linear systems [6,7] as well as in the construction of the inverse of a linear system [8,9]. An analogous condition and a recursive relation are not known for the case of discrete-time linear systems.

The present paper presents such a simplified sufficient condition for controllability of discrete-time-varying linear systems and an associated recursive difference relation.

II. PRELIMINARIES

Discrete-time systems are the natural mathematical models for digital

---

*Bracketed numbers indicate references listed at the end of the paper.

PROCEEDINGS, Seventh Annual Allerton Conference on Circuit and System Theory, University of Illinois, October 1969.

systems and, in most cases, they provide an adequate description of sampled-data systems. A one-input one-output discrete-time system may be described by a difference equation. However, this as well as the many-input many-output case can be better described by a set of state difference equations. This description is adopted here.

Let a discrete-time-varying system be described by the state difference equations

$$\underline{q}(k+1) = A(k)\underline{q}(k) + B(k)\underline{x}(k) \tag{1a}$$

$$\underline{y}(k) = C(k)\underline{q}(k) + D(k)\underline{x}(k) \tag{1b}$$

where the n-vector $\underline{q}(k)$ is the state,
the r-vector $\underline{x}(k)$ is the input,
the m-vector $\underline{y}(k)$ is the output,
the matrices $A(k)$,$B(k)$,$C(k)$ and $D(k)$ are of size nxn,nxr,mxn,mxr, respectively and known for all k, and
k is the integer valued discrete time.

If the state $\underline{q}(k)$ at the initial time k_o is $\underline{q}(k_o)$, then for any $k > k_o$ [10]

$$\underline{q}(k) = \Phi(k,k_o)\underline{q}(k_o) + \sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)\underline{x}(i) \tag{2}$$

where by definition, the fundamental matrix of the system is

$$\Phi(k,k') \triangleq \begin{cases} I \text{ (identity matrix) for } k = k' \\ \prod_{j=k'}^{k-1} A(j) = A(k-1)A(k-2)\ldots A(k') \text{ for } k > k' \end{cases} \tag{3}$$

III. CONTROLLABILITY

Definition 1. A discrete-time system is said to be completely controllable if for any initial time k_o and some finite time $k > k_o$ there exists a sequence of inputs $[\underline{x}(k_o),\ldots,\underline{x}(k)]$ which transfers any initial state $\underline{q}(k_o)$ to any specified value $\underline{q}^*$, i.e., it makes $\underline{q}(k) = \underline{q}^*$.
It is noted that the ability to transfer any initial state to any final state in the context of definition 1, is equivalent to being able to reach any final state from the origin. To clarify this statement rewrite Eq. (2) as

$$\underline{q}(k) - \Phi(k,k_o)\underline{q}(k_o) = \sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)\underline{x}(i), \quad k > k_o \tag{4}$$

and define

$$\underline{q}^d \triangleq \underline{q}^* - \Phi(k,k_o)\underline{q}(k_o) . \tag{5}$$

The state $\underline{q}^d$ is completely arbitrary since $\underline{q}^*$ can be any arbitrary value. Now, $\underline{q}(k)$ can be made equal to any $\underline{q}^*$ if and only if the vector $\underline{q}(k) - \Phi(k,k_o)\underline{q}(k_o)$ can be transferred from its zero initial value to any arbitrary $\underline{q}^d$ (corresponding to $\underline{q}^*$) in some finite time k, i.e., if and only if there exists a sequence $[\underline{x}(k_o),\ldots,\underline{x}(k)]$ such that the vector

$$\underline{z}(k) = \sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)\underline{x}(i) \tag{6}$$

can be made equal to any arbitrary value $\underline{q}^d$. Consider the projection of $\underline{z}(k)$ on a nonzero vector $\underline{\lambda}$, i.e.

$$\underline{\lambda}\cdot\underline{z}(k) = \underline{\lambda}^T\underline{z}(k) = \sum_{i=1}^{n} \lambda_i z_i(k) \tag{7}$$

Then, if for every $\underline{\lambda} \neq \underline{0}$ the projection of $\underline{z}(k)$ on $\underline{\lambda}$ can be set to any arbitrary value, the vector $\underline{z}(k)$ can be made equal to any desired value $\underline{q}^d$, i.e. $\underline{z}(k)$ can be set arbitrarily in any direction. Note that

$$\underline{\lambda}^T\underline{z}(k) = \underline{\lambda}^T \sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)\underline{x}(i)$$

$$= \sum_{i=k_o}^{k-1} \underline{\lambda}^T\Phi(k,i+1)B(i)\underline{x}(i) . \tag{8}$$

Choose the input

$$\underline{x}(i) = \alpha B^T(i)\Phi^T(k,i+1)\underline{\lambda} \tag{9}$$

with α a nonzero arbitrary real scalar. Use Equations (8) and (9) to obtain

$$\underline{\lambda}^T\underline{z}(k) = \sum_{i=k_o}^{k-1} \underline{\lambda}^T\Phi(k,i+1)B(i)\alpha B^T(i)\Phi^T(k,i+1)\underline{\lambda} \tag{10}$$

$$= \alpha \sum_{i=k_o}^{k-1} [B^T(i)\Phi^T(k,i+1)\underline{\lambda}]^T[B^T(i)\Phi^T(k,i+1)\underline{\lambda}] \tag{11}$$

Since $\alpha \neq 0$, the projection $\underline{\lambda}^T\underline{z}(k)$ is not zero if and only if

$$\sum_{i=k_o}^{k-1} [B^T(i)\Phi^T(k,i+1)\underline{\lambda}]^T[B^T(i)\Phi^T(k,i+1)\underline{\lambda}] \neq 0 \tag{12}$$

i.e., if and only if the vector

$$\underline{v}(i) \triangleq B^T(i)\Phi^T(k,i+1)\underline{\lambda} \tag{13}$$

is not zero for some i in the set $[k_o,\ldots,k-1]$; then the value of the projection of $\underline{z}(k)$ in the direction of $\underline{\lambda}$ can be adjusted through the choice of the scalar α.

If, on the other hand, $\underline{v}(i)$ is identically zero for all i in the set, then the sum is zero and so is the projection of $\underline{z}(k)$ on $\underline{\lambda}$. Hence the

Theorem 1. A discrete-time system is completely controllable if and only if

given any initial time k_o, there exists a finite time $k > k_o$ such that for every arbitrary vector $\underline{\lambda} \neq \underline{0}$, the vector $\underline{v}(i) = B^T(i)\Phi^T(k,i+1)\underline{\lambda}$ is not zero for some i in the set $[k_o,\ldots,k-1]$.
The proof of the converse is straightforward.

Equation (10) can be rewritten as

$$\underline{\lambda}^T\underline{z}(k) = \underline{\lambda}^T\Big[\sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)B^T(i)\Phi^T(k,i+1)\Big]\underline{\lambda} \tag{14}$$

or

$$\underline{\lambda}^T\underline{z}(k) = \underline{\lambda}^T P(k,k_o)\underline{\lambda} \tag{15}$$

where by definition

$$P(k,k_o) \triangleq \sum_{i=k_o}^{k-1} \Phi(k,i+1)B(i)B^T(i)\Phi^T(k,i+1) \tag{16}$$

The form of (16) suggests that $P(k,k_o)$ is at least positive semi-definite (if not positive definite). In addition, controllability requires that $P(k,k_o)$ has no zero eigenvalues so that $\underline{z}(k)$ may have nonzero projection on every nonzero $\underline{\lambda}$ according to Equation (15). Therefore, $P(k,k_o)$ is nonsingular and hence it is positive definite.

Theorem 2. The system (1) is controllable if and only if given any k_o, the matrix $P(k,k_o)$ is nonsingular for some finite value of $k > k_o$.

The converse can be proved easily.

According to Theorem 1, a necessary and sufficient condition for controllability is that

$$\underline{\lambda}^T\Phi(k,i+1)B(i) \neq \underline{0}^T \tag{17}$$

for any k_o, some finite $k > k_o$, some i in the set $[k_o,\ldots,k-1]$, and every $\underline{\lambda} \neq \underline{0}$. This condition can be used to yield another theorem. Equivalently, condition (17) states that some of the following conditions are simultaneously satisfied for some $i \geq k_o$, some $k > i$, and every $\underline{\lambda} \neq \underline{0}$.

$$\begin{aligned}
&\underline{\lambda}^T\Phi(i+1,i+1)B(i) \neq \underline{0}^T\\
&\underline{\lambda}^T\Phi(i+2,i+1)B(i) \neq \underline{0}^T\\
&\underline{\lambda}^T\Phi(i+2,i+2)B(i+1) \neq \underline{0}^T\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\underline{\lambda}^T\Phi(k,i+1)B(i) \neq \underline{0}^T\\
&\underline{\lambda}^T\Phi(k,i+2)B(i+1) \neq \underline{0}^T\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\underline{\lambda}^T\Phi(k,k)B(k-1) \neq \underline{0}^T
\end{aligned} \tag{18}$$

or, after introducing (3),

$$\underline{\lambda}^T[B(i) \vdots A(i+1)B(i) \ , \ B(i+1) \vdots A(i+2)A(i+1)B(i) \ ,$$

$$A(i+2)B(i+1), \ B(i+2) \vdots \ldots \vdots A(k-1)\ldots A(i+1)B(i) \ ,\ldots,B(k-1)] \neq \underline{0}^T \tag{19}$$

For condition (19) to be satisfied for every $\underline{\lambda} \neq \underline{0}$ it must be true that

$$\text{rank}[B(i) \vdots \ldots \text{same as in (19)} \ldots, \ B(k-1)] = n \tag{20}$$

for any k_o, some $i \geq k_o$, and some $k > i$.

In particular, if Equation (20) holds true for $k = i+n$ then the system is controllable. Now matrices $M_j(i)$ can be defined such that

$$M_o(i) = B(i)$$

and (21)

$$M_j(i) = \mathcal{E} M_{j-1}(i) + A(i+j)M_{j-1}(i)$$

where $\mathcal{E}$ is the "advance" operator, i.e. $\mathcal{E} w(i) = w(i+1)$, and $j = 1,2,\ldots$ Then consider the matrix

$$\Gamma_n(i) \triangleq [M_o(i) \vdots M_1(i) \vdots \ldots \vdots M_{n-1}(i)] \tag{22}$$

and note (after performing some algebra) that every block $M_j(i)$ of $\Gamma_n(i)$ is a linear combination of the subblocks of the jth major block of the matrix in Equation (20) with $k = i+n$. Therefore, if

$$\text{rank}[M_o(i) \vdots M_1(i) \vdots \ldots \vdots M_{n-1}(i)] = n \tag{23}$$

for any k_o and some $i \geq k_o$, then Equation (20) and conditions (19),(18), and (17) are satisfied and the system is controllable. Hence the

Theorem 3. If for any k_o and some $i \geq k_o$ the rank of $\Gamma_n(i)$ is equal to n, then the system (1) is controllable. (A sufficient condition for controllability.)

Since Theorem 3 must be satisfied for any k_o no matter how large, one can have the simpler but stronger condition of

Theorem 4. If the rank of $\Gamma_n(i)$ does not remain less than n for all i as $i \to \infty$, then the system (1) is controllable.

It is noted that expressions (21) and (22) as well as condition (23) have continuous-time counterparts [5]. Also if the system (1) is time-invariant, i.e. A(k) and B(k) are fixed, then condition (3) reduces to a form which is equivalent to the well-known necessary and sufficient condition for controllability of time-invariant systems

$$\text{rank}[B \vdots AB \vdots \ldots \vdots A^{n-1}B] = n. \tag{24}$$

IV. EXAMPLES

The following simple examples illustrate some important points of the

preceding section.

The system shown in Fig. 1 is described by the difference equations

$$\begin{bmatrix} q_1(k+1) \\ q_2(k+1) \end{bmatrix} = \begin{bmatrix} a_1(k) & 0 \\ 0 & a_2(k) \end{bmatrix} \begin{bmatrix} q_1(k) \\ q_2(k) \end{bmatrix} + \begin{bmatrix} b_1(k) \\ b_2(k) \end{bmatrix} x(k) \tag{25}$$

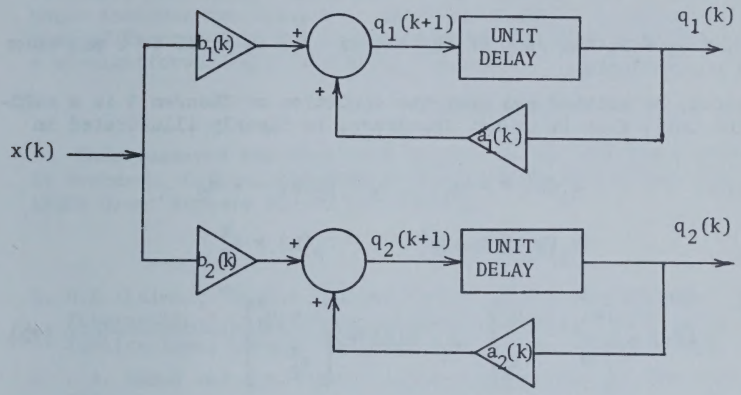


Figure 1

Example 1. Let

$$a_1(k) = -a_1 \quad , \quad a_2(k) = -a_2$$
$$b_1(k) = a_1^k \quad , \quad b_2(k) = a_2^k$$

then

$$A(k) = \begin{bmatrix} -a_1 & 0 \\ 0 & -a_2 \end{bmatrix} \quad , \quad B(k) = \begin{bmatrix} a_1^k \\ a_2^k \end{bmatrix} . \tag{26}$$

Equations (21) and (22) yield,

$$M_o(i) = B(i) = \begin{bmatrix} a_1^i \\ a_2^i \end{bmatrix} \tag{27}$$

$$M_1(i) = \begin{bmatrix} a_1^{i+1} \\ a_2^{i+1} \end{bmatrix} + \begin{bmatrix} -a_1 & 0 \\ 0 & -a_2 \end{bmatrix} \begin{bmatrix} a_1^i \\ a_2^i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{28}$$

$$\Gamma_2(i) = \begin{bmatrix} a_1^i & \vdots & 0 \\ \cdots & \cdots & \cdots \\ a_2^i & \vdots & 0 \end{bmatrix} \tag{29}$$

and

$$\text{rank } \Gamma_2(i) = 1 \tag{30}$$

which is less than 2, the order of the system. Therefore Theorem 3 is not applicable and yields no information concerning the controllability of the system under consideration.

However, it is possible to examine the rank of the matrix of condition (19); this yields

$$[B(i)\vdots A(i+1)B(i),\ B(i+1)] = \begin{bmatrix} a_1^i & \vdots & -a_1^{i+1} & \vdots & a_1^{i+1} \\ \vdots & & \vdots & & \vdots \\ a_2^i & \vdots & -a_2^{i+1} & \vdots & a_2^{i+1} \end{bmatrix} \tag{31}$$

Clearly for $a_1 \neq a_2 \neq 1$, the rank of the matrix (31) is equal to 2 and hence the system is controllable.

Here it should be pointed out that the condition of Theorem 3 is a sufficient condition only; that it is not necessary, is clearly illustrated in this example.

Example 2. Let

$$a_1(k) = -a_1 \quad , \quad a_2(k) = -a_2$$

$$b_1(k) = 1 + a_1^k \quad , \quad b_2(k) = a_2^k$$

then

$$A(k) = \begin{bmatrix} -a_1 & 0 \\ 0 & -a_2 \end{bmatrix} \quad , \quad B(k) = \begin{bmatrix} 1 + a_1^k \\ a_2^k \end{bmatrix}. \tag{32}$$

Again Eqs. (21) and (22) yield,

$$M_o(i) = B(i) = \begin{bmatrix} 1 + a_1^i \\ a_2^i \end{bmatrix} \tag{33}$$

$$M_1(i) = \begin{bmatrix} 1 + a_1^{i+1} \\ a_2^{i+1} \end{bmatrix} + \begin{bmatrix} -a_1 & 0 \\ 0 & -a_2 \end{bmatrix} \begin{bmatrix} 1 + a_1^i \\ a_2^i \end{bmatrix} = \begin{bmatrix} 1 - a_1 \\ 0 \end{bmatrix} \tag{34}$$

and

$$\Gamma_2(i) = \begin{bmatrix} 1 + a_1^i & \vdots & 1 - a_1 \\ & \vdots & \\ a_2^i & \vdots & 0 \end{bmatrix} \tag{35}$$

Provided that $a_1 \neq 1$, the rank of $\Gamma_2(i)$ is equal to 2 (the order of the system) for all i and hence the system is controllable according to Theorem 3. Theorem 4 is also applicable in this case; in particular, if one assumes $|a_1|$ and $|a_2| < 1$, then

$$\lim_{i\to\infty} \Gamma_n(i) = \lim_{\varepsilon\to 0} \begin{bmatrix} 1 & 1-a \\ \varepsilon & 0 \end{bmatrix} \tag{36}$$

and the rank of $\Gamma_n(i)$ does not remain less than n as $i\to\infty$; hence the system is controllable.

V. CONCLUSIONS

This paper presents two sufficient conditions for controllability of discrete-time-varying systems; they are given as Theorems 3 and 4. The known Theorems 1 and 2 are included for the sake of completeness and in order to facilitate the proof of Theorems 3 and 4. It is safe to say that Theorem 3 and the recursive relation (21), which is associated with it, should play the same central role as their continuous-time counterparts do, in the solution of several important problems in the theory of linear systems. For time-invariant systems, Theorem 3 yields the well known necessary and sufficient matrix condition for controllability of such systems. The paper includes two examples. The first example illustrates that the condition of Theorem 3 is not a necessary one for controllability; the second is a straightforward application of Theorem 4.

ACKNOWLEDGMENT

This research was sponsored in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant Numbers 499-67 and 68-1393.

REFERENCES

1. R.E. Kalman, "On the General Theory of Control Systems," Proc. First International Congress on Automatic Control, Butterworths Scientific Publications, London, England, Vol. 1, pp. 481-492, 1961.
2. L.A. Zadeh and C.A. Desoer, Linear System Theory, New York: McGraw-Hill, 1963.
3. E. Kreindler and P.E. Sarachik, "On the oncepts of Controllability and Observability of Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-9, pp. 129-136, April 1964.
4. P.E. Sarachik and E. Kreindler, "Controllability and Observability of Linear Discrete-Time Systems," International Journal of Control, Vol. 1, pp. 419-432, May 1965.
5. L.M. Silverman and H.E. Meadows, "Controllability and Time-Variable Unilateral Networks," IEEE Trans. on Circuit Theory, Vol. CT-12, pp. 308-314, September 1965.
6. L.M. Silverman and H.E. Meadows, "Degrees of Controllability in Time-Variable Linear Systems," Proc. of the National Electronic Conference, Vol. XXI, pp. 689-693, 1965.
7. L. Weiss and P.L. Falb, "Doležal's Theorem, Linear Algebra with Continuously Parametrized Elements, and Time-Varying Systems," Mathematical Systems Theory, Vol. 3, pp. 67-75, 1969.
8. L. M. Silverman, "Properties and Application of Inverse Systems," IEEE Trans. on Automatic Control, Vol. AC-13, pp. 436-437, August 1968.
9. L.M. Silverman, "Inversion of Multivariable Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-14, pp. 270-276, June 1969.
10. R.J. Schwarz and B. Friedland, Linear Systems, New York: McGraw-Hill, 1965.

A SAMPLING METHOD FOR IDENTIFICATION OF LINEAR CONTINUOUS SYSTEMS

I. ELDEM
Lieutenant, Turkish Navy
Naval Postgraduate School
Monterey, California

G. J. THALER
Professor
Naval Postgraduate School
Monterey, California

A continuous, linear, stationary system is representable in state variable form by

$$\dot{\underline{x}} = A\underline{x} + B\underline{u} \tag{1}$$

$$\underline{Y} = C\underline{x}$$

and in like manner a discrete, linear, stationary system is representable by

$$\underline{x}(k+1) = A_D\underline{x}(k) + B_D\underline{u}(k)$$

$$\underline{Y}(k) = C_D\underline{x}(k) \tag{2}$$

Ho's [1] algorithm provides a method for determining the A, B, C (or A_D, B_D, C_D) matrices.

For discrete systems matrices are formed using output samples at successive sampling instants, and these are manipulated according to the algorithm to obtain the state model. For continuous systems the algorithm follows the same pattern except that the sample values must be replaced by the impulse response and its successive derivatives evaluated at t = 0.

Application of these ideas (using digital computer simulation) shows that for discrete systems the identification is exactly correct; i.e., the only limitation is the accuracy with which the samples are measured. For continuous systems the results are not satisfactory unless great care is taken and the measurements are very accurate.

The main error source in the above procedure is the evaluation of derivatives numerically. If the measured data can be used directly, without evaluating derivatives, inclusion of this error to the algorithm can be eliminated.

This approach leads to a system which is operating continuously but observed discretely. If the system has the dynamics:

$$\dot{\underset{\sim}{x}} = A\underset{\sim}{x} + Bu$$

$$Y = C\underset{\sim}{x} \tag{3}$$

Then

$$\underset{\sim}{x} = \mathcal{L}^{-1}\left\{(sI-A)^{-1}\underset{\sim}{x}(t_o) + (sI-A)^{-1}Bu(s)\right\} \tag{4}$$

As u(t) is a step (amplitude V)

$$\underset{\sim}{x}(t) = \mathcal{L}^{-1}\left\{(sI-A)^{-1}\underset{\sim}{x}(t_o) + \frac{(sI-A)^{-1}B}{s} V\right\} \tag{5}$$

Defining

$$\theta(t-t_o) \triangleq \mathcal{L}^{-1}\left\{(sI-A)^{-1}\right\}$$

$$\Gamma(t-t_o) = \mathcal{L}^{-1}\left\{\frac{(sI-A)^{-1}B}{s}\right\} \tag{6}$$

Substituting

$$\underset{\sim}{x}(t) = \phi(t-t_o)\,\underset{\sim}{x}(0) + \Gamma(t-t_o)V \tag{7}$$

letting

$$t_o = kT \tag{8}$$

$$t = t_o + T$$

$$\underset{\sim}{x}[(k+1)T] = \phi(T)\underset{\sim}{x}(kT) + \Gamma(T)V \tag{9}$$

and

$$Y(kT) = C\underset{\sim}{x}(kT) \tag{10}$$

$\phi(T)$ and $\Gamma(T)$ are constant matrices as the sampling period is constant, V can be replaced by v(kT) letting

$$v(kT) = 1,\ k = 0,1,2... \tag{11}$$

as the test input is a unit step.

Dynamics described by (9) and (10) can be represented as a discrete system.

If the continuous system is driven by a unit step and the discrete system with a unit impulse train of period T, they both will give the same y(kT). Therefore samples taken as the result of experiment on the continuous system can be used to identify the system as if it is driven by a unit impulse train and the output is measured.

As a result Ho's algorithm can be applied to the identification of linear continuous systems by applying a step input and sampling throughout. Acceptable identifications are obtained when the sample values are measured accurately to four significant digits.

REFERENCE

[1] Ho, B. L. and R. E. Kalman, Effective Construction of Linear State Variable Models from Input/Output Functions, Regelungstechnik, Jahrg. Heft. 12, 1966.

DETERMINATION OF THE STATE TRANSITION MATRIX OF A SAMPLED-DATA CONTROL SYSTEM FROM MEASUREMENTS OF THE INPUT AND OUTPUT STATES

Willie L. McDaniel, Jr. and Jerrel R. Mitchell
Department of Electrical Engineering
Mississippi State University
State College, Mississippi 39762

ABSTRACT

Two methods for estimating the state transition matrix of a sampled-data control system are presented. The methods involve an averaging technique in one case and a dynamic filtering process in the other case. The estimation is considered under "perfect"measurements and "noisy" measurements.

SUMMARY

The input-output relationship of a typical time invariant sampled-data control system, which has the sampler followed by a zero-order hold is

$$\underline{x}[(n+1)T] = \Phi(T)\,\underline{x}(nT) + \int_0^T \Phi(\sigma)\,B\,d\sigma\;\underline{u}(nT) \qquad (1)$$

where
- T is the uniform sampling period
- $\Phi(T)$ is the state transition matrix, m x m
- $\underline{u}(nT)$ is the input vector, sampled and held, r x 1
- $\underline{x}(nT)$ is the state vector, m x 1
- B is an m x r matrix

A more general system description could be written as

$$\underline{x}(n+1) = \Phi(n+1,n)\,\underline{x}(n) + D\,\underline{u}(n) \qquad (2)$$

where D is a m x r matrix.

In some applications it is possible that little is known about $\Phi(n+1, n)$ and D. In fact, in many instances only the dimensions of $\Phi(n+1, n)$ and D are known; furthermore, the input vector, $\underline{u}(n)$, and the state vectors, $\underline{x}(n+1)$ and $\underline{x}(n)$, are only known through measurements in which there is noise. From the knowledge of the dimensions of the Φ and D matrices and the measurements of the input and output vectors, it often becomes desirable to estimate the Φ and D matrices. Two methods that have been given in the literature use the pulse spectral density and quadratic programming.[1] The purpose of this paper is to present two additional methods for estimating the Φ and D matrices of a sampled-data control system.

First, the necessary information that would be needed to calculate the two matrices, if the input and output vectors were known perfectly is derived. Then, two methods of estimating the matrices are presented. Finally, illustrated examples of each method are given.

The key to the estimation process is the augmentation of the Φ matrix by the vector $\underline{d}$ where a single input system is used. Assuming an inverse exists for a matrix of additional state measurements, assumed perfect,

then it is shown that the augmented Φ matrix, Φ_A , is

$$\Phi(n+1,n) = [\underline{x}(n+1),\ x(n+2),\ \ldots,\underline{x}(m+n+2)]\begin{bmatrix}\underline{x}(n),\ x(n+1),\ldots,\underline{x}(m+n+1)\\ u(n),\ u(n+1),\ldots,u(m+n+1)\end{bmatrix}^{-1} \tag{3}$$

With this as background, two methods of determining Φ_A are presented. The first method assumes a measurement equation of the form

$$\theta(n) = \underline{x}(n) + \underline{M}(n) \tag{4}$$

with the input equation

$$y(n) = u(n) + N(n) \tag{5}$$

where $\underline{M}(n)$ and $N(n)$ are specified noises with certain statistical properties given. An averaging technique gives the estimate of Φ as

$$\hat{\Phi} = \overline{\underline{M}}_1 \begin{bmatrix}\overline{\underline{M}}_0\\ \overline{N}\end{bmatrix}^{-1}$$

where $\overline{\underline{M}}_1 = [\underline{M}(n+1)\ \underline{M}(n+2)\ \ldots\ \underline{M}(n+m+2)]$, $\overline{\underline{M}}_0 = [\underline{M}(n)\ \underline{M}(n+1),\ldots\underline{M}(n+m+1)]$, and $\overline{N} = [N(n)\ N(n+1)\ \ldots\ N(n+m+1)]$.

Using a Kalman estimation procedure,[2] the Φ matrix can be estimated through an iteration procedure described by a digital program given in Reference 3.

Examples illustrate methods, but point up that the technique might not be applicable under all circumstances.

REFERENCES

1. Willet, Richard M., "Sampled-Data Adaptive Control Utilizing Variable Sampling Rates", Ph.D. Dissertation, Iowa State University of Science and Technology, June 1966.

2. Liebelt, Paul B., The Boeing Company, An Introduction to Optimal Estimation, Addison-Wesley Publishing Company Reading, Mass., 1967.

3. McDaniel, Willie L. and Mitchell, Jerrel R., "Computational Schemes in Estimation Problems," Graduate Seminar Program, Electrical Engineering Department, Mississippi State University, Fall 1968.

A MAXIMUM PRINCIPLE FOR SAMPLED DATA SYSTEMS WITH TRAJECTORY CONSTRAINTS

JUERGEN BRUCKNER
Battelle Memorial Institute
Columbus Laboratories
Columbus, Ohio

SHERMAN WU
Martin Marietta Corporation
Martin Co., Division
Orlando, Florida

ABSTRACT

Trajectory constraints are imposed on a continuous system whose evolution is described by difference equations. A first order approximation to the actual (and unknown) trajectory is used. The state constraints on the approximated trajectory are converted to mixed state and control constraints on a discrete trajectory. Necessary conditions for an optimal solution can then be given.

INTRODUCTION

The requirement on a system described by difference equations may restrict the system's trajectory, $x_0, x_1, \ldots, x_k$ (herein called the discrete trajectory), to lie within a prescribed region. In the past [1] [2] this requirement has assumed the form

$$x_i \in Q = \{x : q(x) \leq 0\} \text{ for every } i = 0, 1, \ldots, k. \tag{1}$$

However, a difficulty arises when the discrete system is an incomplete description of a continuous system due, for example, to limitations on the control and/or trajectory storage. The continuous trajectory is still required to lie outside a region, while the trajectory is given at discrete points in space. Hence, a modified system trajectory, called the indiscrete trajectory, is proposed. The indiscrete trajectory, $x_i(\Delta)$, will assume the ray between x_i and x_{i+1} for $i=0,1,\ldots,k-1$. Formally,

$$x_i(\Delta) = (1-\Delta)x_i + \Delta x_{i+1} \text{ for every } i = 0, 1, \ldots, k-1, \forall \Delta \in [0,1]. \tag{2}$$

The entire indiscrete trajectory will be required to satisfy the inequality constraint of Equation (1).

PROBLEM FORMULATION

The evolution of the system is described by the difference equation

$$x_{i+1} - x_i = f_i(x_i, u_i) \text{ for every } i = 0, 1, \ldots, k-1, \text{ where}$$

$$x_i \in E^n \text{ and } u_i \in U_i \subset E^m . \tag{3}$$

The requirement that the indiscrete trajectory remain in the allowable region can be stated as follows:

$$x_i(\Delta) \in Q \; \forall \Delta \in [0,1) \text{ for every } i = 0, 1, \ldots, k-1; \; x_k \in Q . \tag{4}$$

The initial and terminal states satisfy

$$h(x_o) = 0, \; h: E^n \to E^{n_o}, \text{ and } g(x_k) = 0, \; g: E^n \to E^{n_k} \tag{5}$$

respectively. It is required to find a trajectory sequence $x_0, x_1, \ldots, x_k$, and control sequence $u_0, u_1, \ldots, u_{k-1}$ (called optimal), such that an objective functional is maximized while satisfying the above constraints.

SOLUTION

The necessary conditions for optimality will be stated in terms of an auxiliary function $s(x,u)$ which transforms the requirements on the indiscrete trajectory to requirements on the discrete trajectory and control sequences. Before proceeding, it is convenient to introduce the following terminology. Define ∂Q, the boundary of Q, as follows:

$$\partial Q = \{x : q(x) = 0\} . \tag{6}$$

Any $x \in \partial Q$ is called a boundary point of Q. The trajectory is said to traverse the boundary at time i whenever $x_i(\Delta_i) \in \partial Q$ for $0 \leq \Delta_i < 1$. $x_i(\Delta_i)$ is then called the contact point for x_i.

Initially, assume that $q \in C^{(1)}$. Consider the scalar $r[x_i,u_i,x_i(\Delta_i)]$ defined on ∂Q as follows:

$$r[x_i,u_i,x_i(\Delta_i)] = \frac{\partial q[x_i(\Delta_i)]}{\partial x} f_i(x_i,u_i) \tag{7}$$

Since $x_i(\Delta_i)$ is a boundary point,

$$q[x_i(\Delta_i) + \delta f_i(x_i,u_i)] = q[x_i(\Delta_i)] + \delta \frac{\partial q[x_i(\Delta_i)]}{\partial x} f_i(x_i,u_i) + o(\delta) \tag{8}$$

$$= \delta\, r[x_i,u_i,x_i(\Delta_i)] + o(\delta) \tag{9}$$

$$\leq 0 \text{ for } 0 \leq \Delta_i + \delta \leq 1 , \tag{10}$$

which implies that

$$r[x_i,u_i,x_i(\Delta_i)] \leq 0 \text{ for } \Delta_i = 0, \text{ and} \tag{11}$$

$$r[x_i,u_i,x_i(\Delta_i)] = 0 \text{ for } 0 < \Delta_i < 1 . \tag{12}$$

The trajectory is said to be on the boundary whenever $r = 0$, and leaving the boundary when $r < 0$.

Now assume $r = 0$. Then by the implicit function theorem [3, p 13], there exists a function $X_i(x_i,u_i)$ such that

$$r[x_i,u_i,X_i(x_i,u_i)] = 0 . \tag{13}$$

Define $s(x_i,u_i)$ as follows:

$$s(x_i,u_i) = r[x_i,u_i,X_i(x_i,u_i)] \text{ for } x_i(\Delta_i) \neq x_i;$$

$$= r[x_i,u_i,x_i] \text{ for } x_i(\Delta_i) = x_i . \tag{14}$$

Then, whenever the indiscrete trajectory is on the boundary, the discrete trajectory must satisfy the constraint

$$s(x_i,u_i) = 0 . \tag{15}$$

In order to define $s(x_i,u_i)$ over the entire trajectory, let $s(x_i,u_i) = q(x_i)$ when x_i has no contact point, i=0,1,...,k-1 and $s(x_k,u_k) = q(x_k)$. Then by requiring $s \in C^{(1)}$ the variational principle of Pearson and Sridhar in [2], or by further requiring $s \in C^{(2)}$, the maximum principle developed by Bruckner in [4], may be used (after satisfying certain continuity and rank requirements on the remaining defining equations) to obtain the optimal solution.

REFERENCES

[1] Canon, M., Cullum, C., and Polak, E., "Constrained Minimization Problems in Finite Dimensional Spaces", J. SIAM Control, Vol 4, No. 3, pp 528-547 (1966).

[2] Pearson, J. B., and Sridhar, R., "A Discrete Optimal Control Problem", IEEE Trans. on Automatic Control, Vol AC-11, No. 2, pp 171-174 (1966).

[3] Caratheodory, C., Calculus of Variations and Partial Differential Equations of the First Order, Part 1: (R. B. Dean and J. J. Brandstatter, trans.), Holden-Day, Inc., San Francisco (1965).

[4] Bruckner, J., Optimal Control of Discrete Systems, Ph.D. dissertation, Marquette University, Milwaukee (1969).

ON QUANTIZED CONTROL SIGNAL SYSTEMS

R. MARK HAVIRA AND JOHN B. LEWIS
Pennsylvania State University
University Park, Pennsylvania

Abstract - The optimal control of continuous-time quantized control signal systems has not been studied extensively beyond the cases of two and three level controls. Here a design method is presented for continuous-time control systems in which the plant control signal is quantized to 2, 3, 5,..., levels. The control signal is assumed to be constrained in magnitude. The Minimum Principle is then used. A new cost functional is introduced, and this cost functional is such that it includes minimum-time and minimum-fuel problems. An example is then given in which there are seven levels of control.

INTRODUCTION

The design of linear and nonlinear control systems has progressed to the point where general synthesis procedures now exist. For nonlinear systems, particularly in the area of continuous-time systems, the largest amount of work has been directed toward studies of minimum-time systems--the time-optimal or "bang-bang" problem--and the minimum-fuel problem (1). The optimal control for these two problems can be implemented by use of relays and relatively simple switch functions which makes them attractive for practical reasons. However, the conditions under which these systems operate must be ideal, that is there must be no errors in switching between levels if optimal performance is to result. In fact, inaccuracies in the plant model or the feedback signals do exist due to sensitivity of system parameters to environmental influences or external disturbances. It is this sensitivity which makes the possibility of multi-level or quantized systems more attractive than two level systems.

Previous studies of optimal multi-level control systems have only been done for discrete-time systems by Lewis (4), Lewis and Tou (5), and Larson (3). The mathematical method used was dynamic programming or a modified dynamic programming approach. An investigation of related continuous-time control systems was done by Havira (2) using the Minimum Principle. Here a more general formulation is presented and the Minimum Principle is used since it can take care of a large class of constraints.

PROBLEM STATEMENT

The mathematical description of the general optimal quantized problem is presented here and a new cost functional for the control force is introduced. The cost functional is such that minimum-time and minimum-fuel problems are included in the formulation of the general problem. That is, the cost functional Q(u) is the integral of a function L(u) which is composed of straight line segments. L(u) approximates functions of the form

$$L^S(u) = |u|^\alpha \qquad \alpha = 0,1,2,---$$

For α=0 and α=1, the cost functional is that for minimum-time and minimum-fuel problems respectively (1). If α>1 the $L^S(u)$ can be approximated by as many straight line segments as desired. A typical cost function L(u) is shown in Figure 1 where the break points m_i are assumed to be known a priori. It is also assumed that the control force is bounded

$$|u| \leq m_N = M \tag{1}$$

where N is one half the total number of straight line segments, and a symmetric L(u) is considered here although many variations are possible. The C_i's are the slopes of L(u) and the general form of L(u) is

$$L(u) = C_i|u| - d_i \qquad \text{if } m_i < |u| \leq m_{i+1} \tag{2}$$

where $i = 0,1,\ldots,N-1$

$m_o = d_o = 0$ and $m_N = M$

For the scalar input case, Q(u) can be defined as

$$Q(u) = \int_{t_o}^{T} L(u)\,dt \tag{3}$$

and for the vector input case as

$$Q = \int_{t_o}^{T} \left\{ \sum_{j=1}^{r} a_j L_j(u_j) \right\} dt \qquad a_j > 0 \tag{4}$$

where the u_j are the components of the control vector $\underline{u}$, L_j is the cost functional for that component, and the a_j are constants; Equation (4) is a general cost functional. In order to simplify the algebra the constants a_j are set equal to one.

The system to be investigated is described by the differ-

ential equation

$$\dot{\underline{x}}(t) = A\underline{x}(t) + B\underline{u}(t) \tag{5}$$

where the n vector $\underline{x}(t)$ is the state, the nxn constant matrix A is the system matrix, the nxr constant matrix B is the drive matrix, and the r vector $\underline{u}(t)$ is the control. It is assumed that the system (5) is completely controllable and the components of $\underline{u}(t)$ are constrained in magnitude by the relations

$$|u_j| \leq 1 \qquad j = 1,2,\ldots,r \qquad \text{for all } t \tag{6}$$

Then, at the initial time $t_o=0$, the known initial state of the system is

$$\underline{x}(0) = \underline{\xi} \tag{7}$$

Given a specified terminal state $\underline{\theta}$ ($\underline{\theta} \neq \underline{\xi}$ and not necessarily $\underline{0}$); given a finite time T_f (It is assumed $T_f > T^* > 0, T^*$ the smallest possible realizable time) then determine the Q-optimal control that transfers the system (5) from $\underline{\xi}$ to $\underline{\theta}$ in time T_f and minimizes the cost functional $Q(\underline{u})$ described in Eq. (4) where a_j's are equal to one.

NECESSARY CONDITIONS

The Hamiltonian for the system (5) and the cost functional $Q(\underline{u})$ is given by the equation

$$H(\underline{x}(t),\underline{p}(t),\underline{u}(t)) = \sum_{j=1}^{r} L_j(u_j(t)) + \langle A\underline{x}(t),\underline{p}(t)\rangle + \langle \underline{u}(t),B^T\underline{p}(t)\rangle. \tag{8}$$

Observe that the Hamiltonian for $\alpha>0$ is linear in $u_j(t)$ and $|u_j(t)|$ since $L_j(u_j)$ has terms $C_{ij}|u_j|+d_{ij}$ for all i. Now if $\underline{u}^*(t)$ is the optimal control and $\underline{x}^*(t)$ the resulting optimal trajectory, in order that $\underline{u}(t)$ be extremal, the costate $\underline{p}^*(t)$ and state $\underline{x}^*(t)$ must be solutions of the canonical equations.

$$\dot{\underline{x}}^*(t) = \frac{\partial H}{\partial \underline{p}} = A\underline{x}^*(t) + B\underline{u}^*(t)$$

$$\dot{\underline{p}}^*(t) = -\frac{\partial H}{\partial \underline{x}} = -A^T\underline{p}^*(t) \tag{9}$$

with the boundary conditions $\underline{x}^*(0) = \underline{\xi}$ and $\underline{x}^*(T_f) = \underline{\theta}$. Furthermore, it is necessary that the Hamiltonian be a positive constant and

$$H(\underline{x}^*(t),\underline{p}^*(t),\underline{u}^*(t)) \leq H(\underline{x}^*(t),\underline{p}^*(t),\underline{u}(t)) \tag{10}$$

for all $|\underline{u}_j(t)| \leq 1$ and $t \in \left[0, T_f\right]$. Then using Eq. (8), Eq. (10) reduces to the relation

$$\sum_{j=1}^{r} L_j(u_j^*(t)) + \langle \underline{u}^*(t), B^T \underline{p}^*(t) \rangle \leq \sum_{j=1}^{r} L_j(u_j(t)) + \langle \underline{u}(t), B^T \underline{p}^*(t) \rangle \tag{11}$$

Let the function $\underline{q}^*(t)$ be defined by the relation

$$\underline{q}^*(t) = B^T \underline{p}^*(t) \tag{12}$$

or equivalently

$$q_j^*(t) = \sum_{k=1}^{n} b_{kj} p_k^*(t) \qquad j=1,2,\ldots,r$$

then Eq. (11) can be written as

$$\sum_{j=1}^{r} (L_j(u_j^*(t)) + u_j^*(t) q_j^*(t)) \leq \sum_{j=1}^{r} (L_j(u_j(t)) + u_j(t) q_j^*(t)) \tag{13}$$

It is claimed that the right side of the Eq. (13) attains its absolute minimum when $\underline{u}(t)$ is as follows

$$\begin{aligned}
u_j(t) &= 0 && \text{if} && |q_j^*| < C_{oj} \\
u_j(t) &= +m_{i+1} && \text{if} && -C_{ij} > q_j^* > -C_{(i+1)j} \\
u_j(t) &= -m_{i+1} && \text{if} && C_{ij} < q_j^* < C_{(i+1)j} \\
m_i \leq u_j(t) &\leq m_{i+1} && \text{if} && q_j^* = -C_{ij} \\
-m_{i+1} \leq u_j(t) &\leq -m_i && \text{if} && q_j^* = +C_{ij} \\
u_j(t) &= -1 && \text{if} && q_j^* > C_{(N-1)j} \\
u_j(t) &= +1 && \text{if} && q_j^* < -C_{(N-1)j}
\end{aligned} \tag{14}$$

This function is shown in Fig. 2. In order to write the above equations in a more compact form, the following notation is used

$$u_j(t) = quz(q_j^*) \tag{15}$$

where quz( ) is defined by Eq. (14). The best way of illustrating the fact that Eq. (14) minimizes the Hamiltonian is to plot the possible values of the function $(L_j(u_j) + u_j q_j^*)$ for all $u_j(t)$ that satisfy the magnitude constraint $|u_j(t)| \leq 1$. This is done in Fig. 3. Depending on the value $u_j(t)$, the value of the function $(L_j(u_j) + u_j q_j^*)$ versus q_j^* will belong to the unshaded area. Note that if q_j^* is $\pm C_{ij}$ the polarity but not the magnitude of $u_j(t)$ is specified. If this is the case over some finite time interval, the problem is singular. It is easy to show that the sufficient conditions for this problem to be normal or the necessary conditions for it to be singular are the same as those for the minimum-fuel problem (1). A similar problem can be formulated where the final time is free. However this is not required since all the necessary conditions are the same as in this problem with the exception of the fact that the Hamiltonian along the optimal trajectory must be zero.

COMPUTED EXAMPLE

Consider the following control problem

$$\begin{aligned} \dot{x}_1 &= x_2 & x_1(0) &= k_1 \\ \dot{x}_2 &= u & x_2(0) &= k_2 \\ |u| &\leq 2 \end{aligned} \tag{16}$$

Minimize

$$Q = \int_0^\infty L(u)dt$$

where L(u) approximates u^2 and is defined as

$$L(u) = \begin{cases} 1/2|u| & \text{if} \quad |u| \leq 1/2 \\ 3/2|u| - 1/2 & \text{if} \quad 1/2 < |u| \leq 1 \\ 3|u| - 2 & \text{if} \quad 1 < |u| \leq 2 \end{cases} \tag{17}$$

while forcing the system for any initial state (k_1, k_2) to the origin (0,0). The control that accomplishes this is given by Eq. (14) and is $u = quz(p_2)$. The control as a function of the state is shown in Fig. 4. Note that for this problem there is

no solution for two regions. The reason for this is that any control which attempts to take the system to (0,0) tends to take it to the x_1 axis. However, once the trajectory reaches the x_1 axis, it will remain there.

CONCLUDING REMARKS

Using the cost functional Q it was shown that the resulting optimal control is quantized with the magnitude of the levels corresponding to the break points of L(u) and the switch times are related to the various slopes of L(u). The Minimum Principle has provided the structure of the control, and a computational method can now be applied to obtain a control law for higher-order systems. Although only a cost was placed on the control, the state also may be included in a more general cost functional $Q(\underline{x}(t),\underline{u}(t))$ and a more general system

$$\dot{\underline{x}}(t) = f(\underline{x}(t),t) + B(\underline{x}(t),t)\underline{u}(t)$$

can be considered.

1. Athans, M. and Falb, P., Optimal Control - An Introduction to the Theory and Its Applications, by McGraw-Hill Book Company, 1966.

2. Havira, R. M., Optimal Control Using Quantized Plant Inputs, MSEE Thesis, Penn. State University, September, 1968.

3. Larson, R. E., "Optimum Quantization in Dynamic Systems", IEEE G-AC Trans., pp. 162-168, April, 1967.

4. Lewis, J. B., Design of Optimum-Quantized-Control-Signal-Sampled-Data Systems, Ph.D. Thesis, Purdue University, August, 1961.

5. Lewis, J. B., and Tou, J. T., "Optimum Sampled-Data Systems with Quantized Control Signals", Trans. AIEE, pp. 195-201, July, 1963.

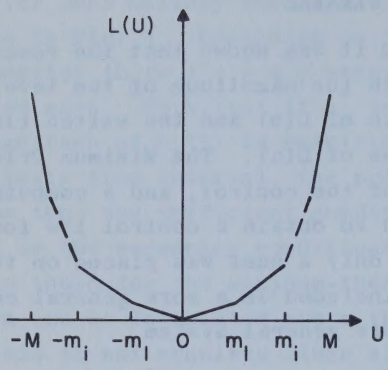


FIGURE 1

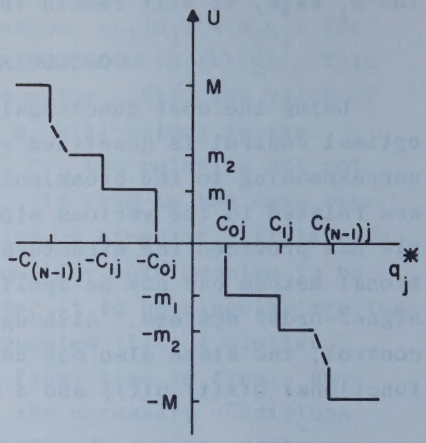


FIGURE 2

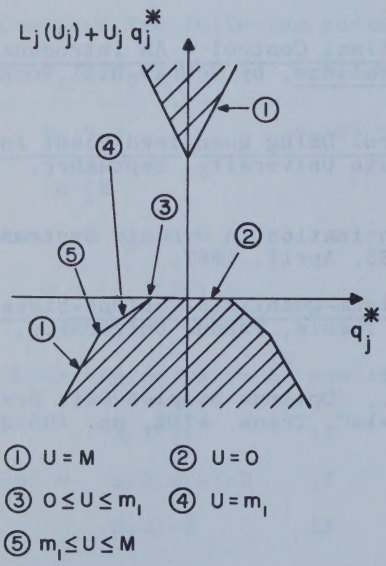


FIGURE 3

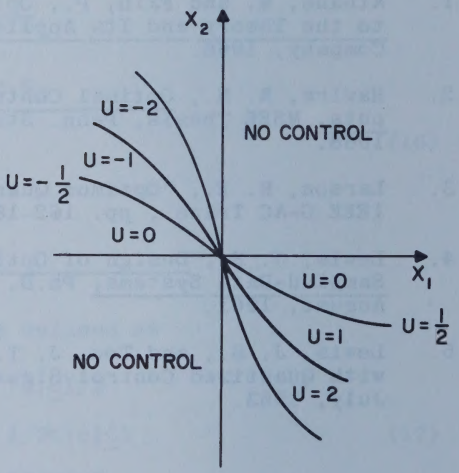


FIGURE 4

FIXED LAG SMOOTHING IN LINEAR DISCRETE SYSTEMS WITH TIME DELAYS

ROLAND PRIEMER
University of Illinois
Chicago, Illinois

ANDRE G. VACROUX
Illinois Institute of Technology
Chicago, Illinois

ABSTRACT

Utilizing the filtering equations derived previously by the authors in the problem of filtering, an algorithm is presented for optimal linear fixed lag smoothing in non-stationary linear discrete systems with multiple time delays.

FIXED LAG SMOOTHING

The fixed lag smoothing filter presented in this paper generates an estimate of the state $x(k-(J+N))$ given the measurements $y(\ell)$, $\ell=0, 1,\ldots,k$, where there are J delay terms in the process described by the vector difference equation:

$$x(k+1) = \sum_{j=0}^{J} \phi_j(k)\, x(k-j) + u(k),$$

and where N=1, 2,..., is the smoothing index, with $N \leq k$ The state x, is an n vector, the $\phi_j(k)$ are time varying $n \times n$ matrices, and u, an n vector, is the random system disturbance. The measurements are modelled by the measurement equation

$$y(k) = D(k)\, x(k) + w(k),$$

where y is an m vector, $m \leq n$ D(k) is a time varying m x n matrix, the rows of which are linearly independent, w, an m vector, is the measurement noise, and k=0, 1,..., is the discrete time index. The random vectors n(k) and w(k) are independent and both are zero mean white Gaussian sequences. The initial states are zero mean Gaussian random vectors, which are independent of u(k) and w(k), for which

$$E\, x(-j)\, x'(-\ell) = V_x(j,\ell)$$

for j, ℓ = 0,..., J, where E denotes the expected value and the prime denotes the matrix transpose.

The method of orthogonal projection was used to derive filtering equations that yield minimum mean square error estimates $\hat{x}(k-j/k)$, j=o,..., J, of the state x(k-j) at the discrete time instant k. Now if the original system is modified to

$$x(k+1) = \sum_{j=0}^{J+N} \phi_j(k)\, x(k-j) + u(k)$$

where: $\phi_j(k) \equiv 0$, for j = J + 1, ..., J + N, then the filtering equations can be applied and become

$$\hat{x}(k-j/k) = \hat{x}(k-j/k-1) + R_j(k)\, \tilde{y}(k/k-1), \quad j = 0,\ldots, J+N$$

$$\tilde{y}(k/k-1) = y(k) - D(k)\, \hat{x}(k/k-1)$$

$$\hat{x}(k+1/k) = \sum_{j=0}^{J} \phi_j(k)\, \hat{x}(k-j/k)$$

where $R_j(k)$, an n x m matrix, is the smoothing gain. As a notational convience, for any integers r, s, and t, define

$$V(r,s/t) = E(x(r) - \hat{x}(r/t))\,(x(s) - \hat{x}(s/t))'.$$

Then the smoothing gain and covariance matrices are given by

$$R_j(k) = V(k-j,\ k/k-1)\ D'(k)\ (D(k)\ V(k,k/k-1)\ D'(k) + V_w(k))^{-1}$$

$$V(k-j,\ k-\ell/k) = V(k-j,\ k-\ell/k-1) - R_j(k)\ D(k)\ V(k,k-\ell/k-1)$$

$$V(k-j,\ k+1/k) = \sum_{\ell=0}^{J} V(k-j,\ k-\ell/k)\ \phi'_\ell(k)$$

$$V(k+1,\ k+1/k) = \sum_{j=0}^{J} \phi_j(k)\ V(k-j,\ k+1/k) + V_u(k)$$

for $j, \ell = 0,\ldots, J,\ldots, J+N$, where V_u and V_w are the system disturbance and measurement noise covariance matrices respectively. Since states prior to $x(-J)$ do not exist, fixed lag smoothing cannot be started until the time $k = N$. Therefore, at $k = 0$, smooth with $N = 0$ using the initial conditions: $V(-j,-\ell/-1) = V_x(j, \ell)$ and $\hat{x}(-j/-1) = 0$, for $j, \ell = 0,\ldots, J$. These equations recursively yield the fixed lag smoothing estimate $\hat{x}(k-(J+N)/k)$ and the smoothing error covariance matrix $V(k-(J+N), k-(J+N)/k)$.

For the particular case of no delay terms, i. e. $J = 0$, the fixed lag smoothing filter reduces to the block diagram shown in the figure.

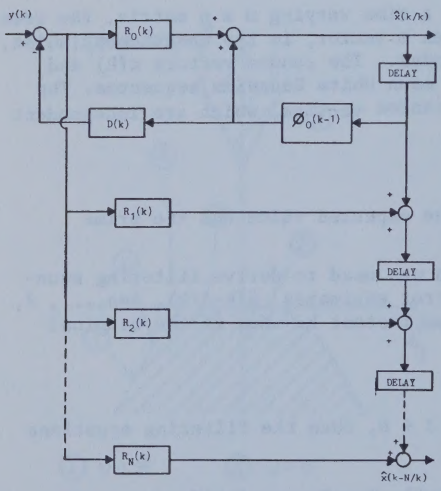


Figure. Optimal Fixed Lag Smoothing Filter.

REFERENCES

R.E. Kalman, "New Approach to Linear Filtering and Prediction Problems," Trans. ASME Ser. D, J. Basic Engrg., 82D (1960) pp. 35-45.

J.S. Meditch, "Orthogonal Projection and Discrete Optimal Linear Smoothing," SIAM J. on Control, Vol. 5, No. 1, 1967, pp. 74-89.

H.E. Rauch, "Solutions to the Linear Smoothing Problem," IEEE Trans. on Automatic Control, AC-8, 1963, pp. 371-372.

R. Priemer and A.G. Vacroux, "Estimation in Linear Discrete Systems with Multiple Time Delays," IEEE Trans. on Automatic Control, Aug., 1969.

R. Priemer and A.G. Vacroux, "Smoothing in Linear Discrete Systems with Multiple Time Delays," Proc. 2nd Asilomar Conf. on Circuits and Systems, 1968.

W.W. Willman, "On the Linear Smoothing Problem," IEEE Trans. on Automatic Control, AC-14, 1969, pp. 116-117.

PROJECTS IN THE HARDWARE RESEARCH GROUP OF THE DEPARTMENT OF COMPUTER SCIENCE OF THE UNIVERSITY OF ILLINOIS

W. J. Poppelbaum

Department of Computer Science
University of Illinois
Urbana, Illinois

I. Industrial Research vs University Research

University research in an area like computer hardware must compete with industrial research. Without admitting this, no successful program can be formulated, i.e. the time of "dreaming alchemists" is definitely gone. String and sealing wax have been replaced by outrageously expensive equipment, and integrated circuits in huge quantities are necessary if prototypes of any scope are to be built. It is, therefore, useful to examine both the similarities and differences between industrial research and university research in order to assess the strategy to be followed in building up a hardware research program.

II. Similarities between Industrial Research and University Research

During the past ten years, industry has done increasingly more basic research -- pure science, as opposed to applied engineering -- which was formerly the province of the university. The university, in order to compete, has had to adopt many of the organizational principles particular to industry. Both industry and the university depend strongly on research funds from government agencies, and this makes the extent of the work that can be performed "seasonal": We are all painfully aware of the consequences of budgetary cuts for the main agencies during the last year. A great similarity also exists in the nature of the research projects because the universities have had to adapt themselves to mission oriented work and strict deadlines, i.e. two factors which up to recently were almost totally missing from academic pursuits. At the same time, industry has adopted certain policies which seemed to be strictly reserved to the university environment, i.e. intensive interaction between in-house research groups and contacts with similar research endeavors elsewhere, even to the extent of granting "sabbaticals" to high ranking industrial personnel.

III. Differences between Industrial and University Research

The most fundamental difference occurs in the staffing of research projects. At the university, the only staff available at a junior level is formed of graduate students without any practical experience. In industry we find trained engineers who already "know the ropes". At the same time, we also have a fundamental difference in the research goals: At the university, research must of necessity lead to M.S. or Ph.D. theses if it is to be useful to its endeavors as an educational institution. In industry, the research goal is new products and ultimately monetary gain. There is another difference, which - slight at first sight - becomes important for the more managerial aspects of the problem and that is the chain of command

difference. In industry, a reasonably clear cut situation exists because of the vertical structure, while the university uses rule by a group of peers. It should be said, even at this point, that the use of young graduate students as researchers also has some outstanding advantages: Because the nature of the selection mechanism for new members of the research group, students of low abilities and enthusiasm are essentially eliminated. This contrasts favorably with the problem of the "retired engineer" in some industrial groups.

IV. Formulation of Some Principles of University Research

Basically the goal must be to turn the negative points of university research mentioned above into assets, while making sure that everybody understands the competitive nature of the operations. This means first of all that projects must be tackled which at the moment seem to be too far out for industry. This has the additional advantage that an enthusiastic graduate student can learn the basics of a proposed new field quicker than his industrial counterpart. At the same time, the classical "know how" of the industrial engineer becomes less important. It should be emphasized that far out projects must be chosen carefully to be realistic enough to produce results: Visible results inside promised deadlines are the best advertisement for a university research group. The horizontal "peer organization" of the university can also be turned into a positive factor, firstly, by using the advice and the facilities of other university research groups and secondly by using the lower inertia of a horizontal structure. Finally, it is essential to implement a project, once it is formulated, with students whose natural inclinations and interests make them particularly suited to the type of work the project demands. In other words, the student's background and his present interests must find a perfect match in the type of project he engages in.

V. Areas of Research of the Hardware Research Group of the Department of Computer Science

The boundary conditions we have mentioned so far have led to a program in four areas such that a modicum of success can be guaranteed, even in the presence of the industrial giants looming in the background. It turns out that practically the success of the group has been so gratifying that industry has, on several occasions, taken from us -- before we were even ready to give up a particular project -- principles which we had developed. This, of course, limits future university work in such an area because of the enormous manpower industry puts into research projects.

The Hardware Research Group in the Department of Computer Science historically designed the circuits for the now dismantled ILLIAC II. With ONR support it then explored ultra-fast circuit techniques and uncommon forms of information processing. For the last five years, with strong support from the Atomic Energy Commission, the group has reached a level of activity characterized by roughly twenty-five graduate students and four senior staff members. The financial resources are sufficient to allow the expenditure of 5 to 10 thousand dollars for a M.S. thesis and 10 to 15 thousand dollars for a Ph.D. thesis. In many cases, a project involves one to two M.S. theses and two to three Ph.D. theses simultaneously.

The Hardware Research Group is presently active in the following four areas:

1. Hybrid Machines, meaning machines which process video signals with up to ten megacycles bandwidth in analog fashion, using digital steering techniques to switch paths.

2. Stochastic Machines, i.e. machines using the principles of the calculus of probabilities to implement numerical processing.

3. Electro-Optical Systems, in which special attention is being paid to interface storage. More precisely, we have specialized in the staticizing in optical form of incoming video information (instant transparencies)!

4. Display Systems, with a strong emphasis on three-dimensional display of computer outputs.

VI. Hybrid Machines

The hybrid machine area has involved us in endeavors as far apart as automatic map coloring and bandwidth compression. The principal projects are:

1. PARAMATRIX - a normalizing picture digitizer which at the same time enhances picture quality and contrast.

2. ARTRIX - a machine which automatically draws line segments indicated by end points and circles indicated by a center and radius, and which more generally corresponds to an electronic drafting table.

3. The Tricolor Cartograph - an interactive color display in which given boundaries lead to the automatic coloration of the "inside", indicated by a light pen.

4. The ORBIT System - in which the information contained in a line drawing is redistributed (for each scanned line) so as to reduce the bandwidth.

5. The VISTA System - in which the scan rate of a television system is automatically adjusted to the rate of change of the incoming information.

VII. Stochastic Machines

The fundamental idea of the stochastic machines is to represent numbers in the range 0 to 1 by the probability of appearance of pulses in fixed time slots. It can easily be seen that the operation of multiplication can then be obtained by a single AND circuit. The other operations of arithmetic can also be implemented quite simply. The remarkable property of these machines is that arbitrarily high precisions can be obtained by sufficiently long integration time and that we have essentially digital circuits processing a form of analog information. The main projects in this area are:

1. POSTCOMP - an infinite precision analog calculator.

2. RASCEL - a tree-type array of stochastic processing elements in which each operating cycle corresponds to the impression of one of the possible arithmetic operations onto each element.

3. TRANSFORMATRIX - a rather large on-line picture processor, in which the intensity of each output point is a linear function of that of all input points and which can, in particular, perform on-line Fourier transforms.

4. APE - a machine in which the interconnections between processing elements can be prescribed and implemented from the outside by the assignment of radio frequency channels.

VIII. Electro-Optical Systems

Realizing fully that the transmission of individual logical signals inside a computer by optical means offers no advantages whatsoever, the main area of work concentrates on systems in which optics are used in their natural environments, i.e. by making the signals to be processed available in parallel optical form on "instant transparencies". The main project is OLFT, a system using a modified Ardenne tube to staticize incoming video signals into a charge distribution. The electro-optic effects in KDP are then called upon to form the spatial Fourier transform of the pattern by means of a laser. Future developments in this area will examine the on-line production of holograms and adaptive filtering in the Fourier transform plane.

A side-project, EROS, examined laser deflection problems in an electronic switching matrix using photo-resistive elements.

IX. Display Systems

The great interest in display systems, especially those leading to three-dimensional displays, is due to the natural enjoyment we obtain when we interact with a computer system in a visual fashion. Presently, two main directions are being followed. In BLAST stereo vision is obtained by writing left eye and right eye information in appropriate places behind a series of vertical cylindrical lenses. In the other, STEREOMATRIX, the computer output is displayed directly by two laser beams polarized at right angles to each other and viewed with appropriate glasses. The system involves variable windowing, a joy stick pointer and ("involuntary") perspective control by the position of the observer.

X. Conclusion

In order to render more immediate the type of research performed in the Hardware Research Group of the University of Illinois, a movie was made in which the principal projects are covered and shown in operation. A reasonably low level explanation of the principles involved makes the movie -- produced by the Motion Picture Services of the University of Illinois -- suitable for public relations purposes.

A RESULT FOR LINEAR SYSTEMS DEFINED OVER ARBITRARY FIELDS: EQUIVALENCE OF CONTROLLABILITY AND POLE ASSIGNMENT*

SANJOY K. MITTER
Electronic Systems Laboratory**
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

1. INTRODUCTION

A result of central importance in the theory of control of autonomous linear systems is the following:

Consider the autonomous linear differential system,

$$\frac{dx}{dt} = A\,x(t) + B\,u(t), \tag{1.1}$$

where A and B are constant $n \times n$ and $n \times m$ matrices respectively, $x(t)$ the state of the system is an element of R^n and the function $t \to u(t)$ is a measurable function defined on $]-\infty,+\infty[$, which is bounded on every compact subinterval of $]-\infty,\infty[$ and takes it values in R^m.

Consider the affine feedback law,

$$u(t) = C\,x(t) + v(t), \tag{1.2}$$

where C is a constant $m \times n$ matrix and $v(t)$ is an external input (satisfying the same hypotheses as $u(t)$).

Let

$$\Lambda = \{\lambda_1,.....,\lambda_n\}\ , \tag{1.3}$$

be an arbitrary set of n complex numbers λ_i, such that any λ_i with $\mathrm{Im}\lambda_i \neq 0$ appears in Λ in a conjugate pair. Assume that the pair (A,B) is completely controllable, that is, the matrix

$$H(A,B) = [B,AB,....,A^{n-1}B] \tag{1.4}$$

has full rank n.

We then have,

Theorem 1.1.

The pair (A,B) is completely controllable if and only if, for every choice of the set Λ, there is a matrix C such that A + BC has precisely Λ for its set of eigenvalues. |

A proof of this result may be found in Wonham [1] and Simon and Mitter [2]. In [2] constructive algorithms to achieve pole-assignment can

---

* This research was partially supported by NSF Grant GK-3714.

**On leave from Case Western Reserve University, Cleveland, Ohio 44106.

be found. These algorithms provide qualitative insight into the problem.

This theorem essentially says that by the application of affine feedback a closed-loop system can be designed to have any arbitrary dynamics, provided the open-loop system is completely controllable. In other words, controllability is equivalent to the property that the closed-loop transfer matrix

$$[sI - (A + BC)^{-1}]B$$

can be assigned an arbitrary set of poles by a suitable choice of the feedback "gain" matrix C. |

This theorem can be generalized to cover linear autonomous finite-dimensional discrete-time linear systems defined over arbitrary fields.

2. NOTATION AND SYSTEM DEFINITION

Let

T = time set = Z = (ordered abelian group of) integers;

U = input values = F^m = vector space of m-tuples over the field F;

X = state space = F^n;

Ω = input space of functions $t \to u(t)$, that is arbitrary sequences $u(-1),u(0),u(1),\ldots$ with $u(t) \in U$;

We shall be concerned with the discrete-time, autonomous, linear dynamical system Σ over a field F,

$$x(t+1) = Ax(t) + Bu(t), \tag{2.1}$$

with $t \in Z$, $x(t) \in F^n$, $u(t) \in F^m$, and where

$$\begin{aligned} A &: F^n \to F^n \\ B &: F^m \to F^n \end{aligned} \tag{2.2}$$

are F-homomorphisms.

We shall usually not make a distinction between (A,B) as a pair of F-homomorphisms or as a pair of matrices representing these homomorphisms with respect to a given basis.

With respect to the system (2.1) we make the assumption: the pair (A,B) is completely controllable, that is, rank of the $n \times nm$ matrix

$$H(A,B) = [B,AB,\ldots,A^{n-1}B] \tag{2.3}$$

is n. |

3. STATEMENT OF MAIN THEOREM

The following theorem can be proved:

Theorem 3.1.

The pair of matrices (A,B) characterizing the discrete-time autonomous system (2.1) over the field F is completely controllable if and only if for every monic polynomial g of degree n, there is a $m \times n$ matrix C over F such that the characteristic polynomial of $A + BC$ is precisely g (up to a factor of -1).

A proof of this theorem can be found in [3].

REFERENCES

1. W. M. Wonham, "On Pole-Assignment in Multi-Input Controllable Linear Systems," IEEE Trans. on Automatic Control, AC-12 (1967), pp. 660-665.

2. J. D. Simon and S. K. Mitter, "A Theory of Modal Control," Information and Control, 13 (1968), pp. 316-353.

3. S. K. Mitter and R. Foulkes, "Controllability and Pole Assignment for Discrete Time Linear Systems Defined Over Arbitrary Fields," to be published in SIAM Journal on Control, 1970.

NOVEL APPLICATIONS OF SCHWARZ'S INEQUALITY

A. PAPOULIS
Polytechnic Institute of Brooklyn
Brooklyn, New York

ABSTRACT

A method is developed for determining the maximum response of a system under various constraints of the input. The solution is based on appropriate extensions of the reasoning leading to the matched filter principle. Special cases include a number of problems treated elsewhere. However, the available solutions involve integral equations and the calculus of variations. In the following approach the results are obtained in simple algebraic terms.

I. INTRODUCTION

Schwarz's inequality states that

$$\left| \int_c^d a(x)b(x)dx \right|^2 \leq \int_c^d |a(x)|^2 dx \int_c^d |b(x)|^2 dx \qquad (1)$$

and that equality holds only if

$$a(x) = k\, b^*(x) \quad . \qquad (2)$$

This simple statement has found extensive applications. In this paper we shall show that by inserting suitable addends or factors in (1) we can use it to solve or simplify a great variety of problems. The discussion will be limited to three typical examples. Other applications are given in reference [1].

II. DRIVING POINT ENERGY CONSTRAINTS

In many applications one is faced with the task of exciting the response $g(t_o)$ of a given system to a specified level using the least possible input energy. This is a special case of the following more general problem: A signal f(t), say a voltage source, is the common input to n given passive networks. The resulting input currents $i_r(t)$ are given by

$$i_r(t) = \int_{-\infty}^{\infty} f(\tau)\, y_r(t-\tau)\, d\tau \qquad r = 1, \ldots, n \qquad (3)$$

where $y_r(t)$ are the corresponding impulse responses. We wish to find f(t) such that the response

$$g(t_o) = \int_{-\infty}^{\infty} f(t)\, h(t_o - \tau)\, d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) H(j\omega) e^{j\omega t_o}\, d\omega \tag{4}$$

is as large as possible and the input energy to each of the n systems does not exceed a specified level:

$$\int_{-\infty}^{\infty} f(t)\, i_r(t)\, dt \leq E_r \,. \tag{5}$$

With

$$Y_r(j\omega) = G_r(\omega) + j\, B_r(\omega) \tag{6}$$

the input admittance, the Fourier transform of the input current is given by

$$I_r(j\omega) = F(j\omega)\, Y_r(j\omega) \tag{7}$$

and the energy constraint (5) takes the form

$$\int_{-\infty}^{\infty} F(j\omega)\, I_r^*(j\omega)\, d\omega = \int_{-\infty}^{\infty} |F(j\omega)|^2\, G_r(\omega)\, d\omega \leq 2\pi E_r \,. \tag{8}$$

From the passivity of the system it follows that

$$G_r(\omega) \geq 0 \,. \tag{9}$$

The function $f_o(t)$ that maximizes $|g(t_o)|$ subject to (5) we shall call optimum.

To solve our problem, we write (9) in the form

$$g(t_o) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) \sqrt{\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)}\; \frac{H(j\omega) e^{j\omega t_o}\, d\omega}{\sqrt{\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)}} \tag{10}$$

where $\alpha_1, \ldots, \alpha_n$ are real constants such that

$$\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega) \geq 0 \,. \tag{11}$$

Applying Schwarz's inequality to the above, we conclude with (8) that

$$|g(t_o)|^2 \leq \sum_{r=1}^{n} \alpha_r E_r \int_{-\infty}^{\infty} \frac{|H(j\omega)|^2 / 2\pi}{\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)}\, d\omega \,. \tag{12}$$

The maximum can be reached only if (12) holds with the equality sign. This is the case if $F(j\omega)$ is such that

$$\sqrt{\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)}\; F(j\omega) = \frac{H^*(j\omega) e^{-j\omega t_o}}{\sqrt{\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)}} \,. \tag{13}$$

We maintain that if a set of n positive constants

$$\bar{\alpha}_r \geq 0 \qquad r = 1, \ldots, n \tag{14}$$

can be found such that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|H(j\omega)|^2\, G_r(\omega)}{[\bar{\alpha}_1 G_1(\omega) + \ldots + \bar{\alpha}_n G_n(\omega)]^2}\, d\omega = E_r \qquad r = 1, \ldots, n \tag{15}$$

then the function

$$F_o(j\omega) = \frac{H^*(j\omega)e^{-j\omega t_o}}{\overline{\alpha}_1 G_1(\omega)+\ldots+\overline{\alpha}_n G_n(\omega)} \tag{16}$$

is the Fourier transform of the optimum signal $f_o(t)$.

Proof. As we see from (15) and (16), the signal $f_o(t)$ satisfies the energy constraint (8). We shall first show that the right side

$$I(\alpha_1,\ldots,\alpha_n) = \sum_{r=1}^{n} \alpha_n E_r \int_{-\infty}^{\infty} \frac{|H(j\omega)|^2/2\pi}{\alpha_1 G_1(\omega)+\ldots+\alpha_n G_n(\omega)}\, d\omega$$

of (12), considered as a function of $\alpha_1,\ldots,\alpha_n$ is minimum if $\alpha_r=\overline{\alpha}_r$, i.e.,

$$I(\alpha_1,\ldots,\alpha_n) \geq I(\overline{\alpha}_1,\ldots,\overline{\alpha}_n)\,. \tag{17}$$

Indeed, if the input to our system is $f_o(t)$ then the corresponding response $g_o(t)$ is such that

$$|g_o(t_o)|^2 \leq I(\alpha_1,\ldots,\alpha_n)$$

[see (12)]. But

$$|g_o(t_o)|^2 = I(\overline{\alpha}_1,\ldots,\overline{\alpha}_n)$$

as we can easily see, hence, (17) is true.

Consider now an arbitrary input f(t) satisfying (5). With g(t) the resulting response and $\gamma_r E_r$ the corresponding input energies, we have

$$0 \leq \gamma_r \leq 1$$

hence, [see (12)]

$$|g(t_o)|^2 \leq \sum_{r=1}^{n} \gamma_r \overline{\alpha}_r E_r \int_{-\infty}^{\infty} \frac{|H(j\omega)|^2 d\omega}{2\pi \sum_{r=1}^{n} \overline{\alpha}_r G_r(\omega)}$$

$$\leq \sum_{r=1}^{n} \overline{\alpha}_r E_r \int_{-\infty}^{\infty} \frac{|H(j\omega)|^2\, d\omega}{2\pi \sum_{r=1}^{n} \overline{\alpha}_r G_r(\omega)} = I(\overline{\alpha}_1,\ldots,\overline{\alpha}_r) = |g_o(t_o)|^2\,.$$

Therefore, $f_o(t)$ is optimum.

We mention without elaboration that if some of the constants $\overline{\alpha}_r$ in (15) are negative but the corresponding sum in (11) is positive, then the constraints (5) of the same subscript are redundant and can be ignored. Solving the problem with the remaining constraints in (5), we obtain an optimum signal that satisfies also the redundant constraints with the inequality sign [1].

III. TIME-LIMITED SIGNALS

In the solution of the main problem of Sec. II the only constraints imposed on the input f(t) were the energy bounds (5). In some cases it is desired that f(t) obeys other restrictions. If we wish f(t) to be band-limited in the band $|\omega| \leq \omega_o$, then we can obtain the optimum input by introducing minor changes in the discussed method. The upper limit of the response $g(t_o)$ is again the right side of (10) where now the integration is over the interval $(-\omega_o, \omega_o)$ and the optimum function $F_o(j\omega)$ is given by (16) for $|\omega| \leq \omega_o$ and it equals zero elsewhere. The constants $\bar{\alpha}_r$ satisfy (15) with the limits of integration replaced by $-\omega_o$ and ω_o. Special cases of the above, involving a single energy constraint are discussed in [2] and [3].

We shall now consider the more difficult problem of a time-limited input [4]. To simplify the notation we shall assume that only one input energy constraint is imposed. The extension to n can be carried out as in Sec. II by replacing $G(\omega)$ by $\alpha_1 G_1(\omega) + \ldots + \alpha_n G_n(\omega)$.

Consider a signal f(t) such that

$$f(t) = 0 \quad \text{for} \quad t < T_1 \quad \text{and} \quad t > T_2 . \tag{18}$$

With f(t) as input voltage to a given system, we assume that

$$\int_{T_1}^{T_2} f(t)\, i(t)\, dt = E , \tag{19}$$

where i(t) is the resulting input current. We wish again to find f(t) such that $|g(t_o)|$ is maximum. The constant T_2 is usually taken equal to t_o, however, it is no more difficult to solve the general problem.

Clearly, the response $|g(t_o)|$ is still bounded by (12), but this bound cannot be reached because the function $F_o(j\omega)$ in (16) is no longer admissible because it does not satisfy (18).

To solve our problem we introduce the functions a(t) and b(t) such that

$$\begin{aligned} a(t) &= 0 \quad \text{for} \quad t > T_1 \\ b(t) &= 0 \quad \text{for} \quad t < T_2 \end{aligned} . \tag{20}$$

With A(p) and B(p) their transforms, it follows from (18), (20), and Parseval's formula that

$$\int_{-\infty}^{\infty} F(j\omega)\, A^*(j\omega)\, d\omega = \int_{-\infty}^{\infty} F(j\omega)\, B^*(j\omega)\, d\omega = 0 . \tag{21}$$

With $G(\omega)$ as in (6), the response $g(t_o)$ can be written in the form

$$g(t_o) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \sqrt{G(\omega)}\, F(j\omega)\, \frac{H(j\omega) e^{j\omega t_o} + A^*(j\omega) + B^*(j\omega)}{\sqrt{G(\omega)}}\, d\omega \tag{22}$$

hence

$$|g(t_o)|^2 \leq \frac{E}{2\pi} \int_{-\infty}^{\infty} \frac{|H(j\omega) e^{j\omega t_o} + A^*(j\omega) + B^*(j\omega)|^2}{G(\omega)}\, d\omega \quad . \tag{23}$$

The maximum can be reached only if $F(j\omega)$ equals the optimum

$$F_o(j\omega) = k\, \frac{H^*(j\omega) e^{-j\omega t_o} + A(j\omega) + B(j\omega)}{G(\omega)} \quad . \tag{24}$$

The above is possible only if this function is admissible i. e., if its inverse $f_o(t)$ is zero outside the interval (T_1, T_2). It is known from the theory of Laplace transforms that [5] this is the case if the function

$$F_o(p) e^{pT_1} \tag{25}$$

is analytic in the right hand p-plane and the function

$$F_o(p) e^{pT_2} \tag{26}$$

is analytic in the left hand p-plane. We shall show that this can be accomplished with the proper choice of the functions A(p) and B(p).

We shall assume, as it is the case for finite order systems, that $H(j\omega)$ and $G(\omega)$ are rational functions of ω. Thus

$$H(p) = \frac{N_o(p)}{D_o(p)} \tag{27}$$

and since $G(\omega)$ is even, it can be written in the form

$$G\left(\frac{p}{j}\right) = \frac{N(p)}{D(p)}\, \frac{N(-p)}{D(-p)} \quad . \tag{28}$$

In (27), $D_o(p)$ has zeros only in the left-hand plane (Hurwitz), and in (28), N(p) and D(p) contain only the left-hand zeros and poles of $G(\omega)$. With

$$A(p) e^{pT_1} = \frac{N_1(p)}{D_1(p)} \qquad B(p) e^{pT_2} = \frac{N_2(p)}{D_2(p)} \tag{29}$$

the first function has poles only in the right-hand plane and the second function only in the left-hand plane [see (20)]. Thus, the functions (25) and (26) take the form (omitting the constant k)

$$\left[\frac{N_o(-p)}{D_o(-p)} e^{p(T_1 - t_o)} + \frac{N_1(p)}{D_1(p)} + \frac{N_2(p)}{D_2(p)} e^{p(T_1 - T_2)}\right] \frac{D(p)}{N(p)} \frac{D(-p)}{N(-p)} \tag{30}$$

and

$$\left[\frac{N_o(-p)}{D_o(-p)} e^{p(T_2 - t_o)} + \frac{N_1(p)}{D_1(p)} e^{p(T_2 - T_1)} + \frac{N_2(p)}{D_2(p)}\right] \frac{D(p)}{N(p)} \frac{D(-p)}{N(-p)} \quad . \tag{31}$$

We wish to choose the polynomials in (29) in such a way that the function (30) is analytic in the right-hand plane and the function (31) analytic in the left-hand plane. This can be done as follows: We set

$$D_1(p) = D_o(-p)D(-p) \qquad D_2(p) = D(p) \; . \tag{32}$$

The denominator of (30) and (31) is now

$$D_o(-p)\,N(p)\,N(-p) \quad .$$

The polynomials $N_1(p)$ and $N_2(p)$ are so chosen that the numerator (30) is zero for every root of the polynomial $D_o(-p)\,N(-p)$ and the numerator of (31) is zero for every root of N(p). The resulting function $F_o(p)$ is thus admissible and the constant k is determined from (19). Details involving the relative degrees of N(p) and D(p) and the availability of sufficient parameters to achieve the stated objective are straightforward and will be omitted.

The above approach avoids the calculus of variations [4] and the resulting integral equations. It is similar to the method used in the solution of the Wiener-Hopf equation [6].

We note that if $t_o < T_2$ then $f_o(t)$ is not zero for $t > t_o$. In this case, the energy delivered to the system in the time interval $t_o < t < T$ is negative, i.e., part of the stored energy in the system is returned to the source.

IV. OPTIMUM SIGNALS WITH SPECIFIED TOTAL AND LOCAL ENERGY

Consider the class C of functions f(t), band-limited by ω_o:

$$F(\omega) = 0 \quad \text{for} \quad |\omega| \geq \omega_o \tag{33}$$

of total energy E_o:

$$E_o = \int_{-\infty}^{\infty} |f(t)|^2 \, dt \tag{34}$$

and specified energy E_1 in the interval (-T, T):

$$E_1 = \int_{-T}^{T} |f(t)|^2 \, dt \quad . \tag{35}$$

Given t_o, we wish to find a function $f_o(t)$ in C such that $|f(t_o)|$ is maximum.

As it is known [5, 7], E_1 must be such that

$$E_1 \leq \lambda_o E_o \tag{36}$$

where λ_o is the maximum eigenvalue of the integral equation

$$\int_{-T}^{T} \frac{\sin \omega_o (t-\tau)}{\pi(t-\tau)} \varphi(\tau) \, d\tau = \lambda \varphi(t) \quad . \tag{37}$$

We assume then that (36) is true and seeks to determine $f_o(t)$ using the preceding methods.

We denote by λ_n and $\varphi_n(t)$ the eigenvalues and eigenfunctions (prolate spheroidal functions) of (37) normalized by

$$\int_{-\infty}^{\infty} \varphi_n^2(t)\, dt = 1 \quad . \tag{38}$$

It then follows that

$$\int_{-T}^{T} \varphi_n^2(t)\, dt = \lambda_n \quad . \tag{39}$$

The function f(t) can be expanded into a series

$$f(t) = \sum_{n=0}^{\infty} a_n \varphi_n(t) \tag{40}$$

where

$$a_n = \frac{1}{\lambda_n} \int_{-T}^{T} f(t)\varphi_n(t)dt = \int_{-\infty}^{\infty} f(t)\varphi_n(t)dt \quad . \tag{41}$$

From (38), (39), and the orthogonality of the functions $\varphi_n(t)$ it follows that

$$E_o = \sum_{n=0}^{\infty} a_n^2 \qquad E_1 = \sum_{n=0}^{\infty} \lambda_n \varphi_n^2(t) \quad . \tag{42}$$

To solve our problem, we write (40) in the form

$$f(t) = \sum_{n=0}^{\infty} \sqrt{\lambda \pm \lambda_n}\ a_n \frac{\varphi_n(t)}{\sqrt{\lambda \pm \lambda_n}} \tag{43}$$

where λ is a real constant. The choice of the sign will be determined presently. Applying the discrete form of Schwarz's inequality to the above and using (42), we obtain

$$|f(t_o)|^2 \leq (\lambda E_o \pm E_1) \sum_{n=0}^{\infty} \frac{\varphi_n^2(t_o)}{\lambda \pm \lambda_n} \quad . \tag{44}$$

Equality holds only if

$$\sqrt{\lambda \pm \lambda_n}\ a_n = k \frac{\varphi_n(t_o)}{\sqrt{\lambda \pm \lambda_n}} \quad . \tag{45}$$

Thus, the optimum function is given by

$$f_o(t) \equiv \Psi(t, t_o; \lambda) = k \sum_{n=0}^{\infty} \frac{\varphi_n(t_o)\varphi_n(t)}{\lambda \pm \lambda_n} \quad . \tag{46}$$

This function is band-limited and the constants k and λ are so chosen as to satisfy (42). We note that

$$|f(t_o)| \leq f_o(t_o) = \frac{\lambda E_o \pm E_1}{k} = k \sum_{n=0}^{\infty} \frac{\varphi_n^2(t_o)}{\lambda \pm \lambda_n} \quad . \tag{47}$$

It is easy to see that the value of λ that minimizes the right side of (44) is such as to satisfy (47).

Inserting (46) into (42), we obtain the system

$$\begin{aligned} E_o &= k^2 \sum_{n=0}^{\infty} \frac{\varphi_n^2(t_o)}{(\lambda \pm \lambda_n)^2} \\ \lambda E_o \pm E_1 &= k^2 \sum_{n=0}^{\infty} \frac{\varphi_n^2(t_o)}{\lambda \pm \lambda_n} \end{aligned} \tag{48}$$

from which the constants k and λ are determined.

The choice of the sign of (48) depends on the value of E_1. With

$$\bar{E} = \frac{E_o \pi}{\omega_o} \int_{-T}^{T} \frac{\sin^2 \omega_o (t-t_o)}{\pi^2 (t-t_o)^2} \, dt \tag{49}$$

the local energy of the kernel of (37), if $E_1 = \bar{E}$, then (48) is satisfied for $\lambda = \infty$. If $E_1 < \bar{E}$, then (48) can be satisfied if the + sign is used, if

$$\bar{E} < E_1 \leq \lambda_o E_o$$

then the - sign must be inserted and the resulting value of λ is greater than λ_o. Finally, if $E_1 > \lambda_o E_o$ then the problem has no solution because the class C is empty.

Optimum signal as resolving kernel

We shall conclude with the following interpretation of the optimum signal $\Psi(t, t_o; \lambda)$: Consider the inhomogeneous integral equation

$$z(t) - \frac{1}{\lambda} \int_{-T}^{T} \frac{\sin \omega_o (t-\tau)}{\pi (t-\tau)} z(\tau) \, d\tau = y(t) \, . \tag{50}$$

As it is known [8], the function z(t) is given by

$$z(t) = y(t) - \frac{1}{\lambda} \int_{-T}^{T} \Gamma(t, \tau; \lambda) \, y(\tau) \, dr$$

where the resolving kernel $\Gamma(t, \tau; \lambda)$ is given by

$$\Gamma(t, \tau; \lambda) = \lambda \sum_{n=0}^{\infty} \frac{\varphi_n(t) \varphi_n(\tau)}{\lambda_n - \lambda}$$

From the above and (46) it follows that

$$\Psi(t, t_o; \lambda) = -\lambda \Gamma(t, t_o; \lambda) \ .$$

Thus, the optimum signal is proportional to the resolving kernel of the inhomogeneous integral equation (50).

REFERENCES

[1] A. Papoulis, "Maximum Response with Input Energy Constraints and the Matched Filter Principle," IEEE Transactions on Circuit Theory, May 1970.

[2] A. Papoulis, Systems and Transforms with Applications in Optics, McGraw-Hill Book Co., 1968.

[3] A. Papoulis, "Limits on Bandlimited Signals," Proc. IEEE, Vol. 55, pp. 1677-1686, October, 1967.

[4] L. A. Beattle, "Minimum Energy Triggering Signals," Proc. IEEE, Vol. 46, pp. 751-757, April, 1958.

[5] A. Papoulis, The Fourier Integral and its Applications, McGraw-Hill Book Co., 1962.

[6] D. C. Youla, "The Solution of a Homogeneous Wiener-Hopf Integral Equation Occurring in the Expansion of Second-order Stationary Random Functions," IRE, PGIT, pp. 187-193, September, 1957.

[7] D. Slepian, H. O. Pollack, and H. T. Landau, Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty Principle, Bell System Tech. J., Vol. 40, No. 1, pp. 49-84, January, 1961.

[8] F. G. Tricomi, "Integral Equations," Interscience Publications, Inc., New York, 1957.

SOME NEW COMPUTATIONAL RESULTS FOR MINIMUM VARIANCE CONTROLLERS

MICHAEL K. SAIN AND STANLEY R. LIBERTY
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

In an earlier paper, one of the authors has given a complete solution to a class of minimum variance control problems. Though established in principle, the solution was difficult to apply because of the need of prior hand calculations, these being straightforward but tedious. As a result of that computational complexity, only the simplest of numerical examples have been exhibited up to this time. In this paper, a new and conceptually much simpler solution is brought forth, and a number of more extensive examples are presented. The new approach makes the minimum variance control concept readily available for numerical experiment.

INTRODUCTION

Consider the problem of controlling the linear, time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t) + \theta(t) \tag{1}$$

according to the quality index

$$\int_0^{t_f} [x'(\sigma)Qx(\sigma) + u'(\sigma)Ru(\sigma)]d\sigma + x'(t_f)Px(t_f) , \tag{2}$$

where (1) is defined on the integration interval of (2) and where t_f is a fixed final time of interest. Control is complicated by the presence of the noise $\theta(t)$, which is assumed to be zero-mean Gaussian with essentially white characteristics within the bandwidth of (1), that is,

$$E\{ \theta(t)\theta'(\sigma) \} = S\delta(t-\sigma) . \tag{3}$$

In (1), state $x(t) \in R^n$ and control action $u(t) \in R^m$. Matrices P, Q, R, S are all symmetric and positive semi-definite, with R also possessing the matrix inverse. The superscript (') denotes transposition. The initial state x(0) is known exactly.

The random variable (2) cannot, of course, be minimized directly. One of the most common methods of optimizing it is to optimize its average value. If the average value of (2) is not minimized, but is merely constrained to a non-minimum value, additional control flexibility becomes available. This additional flexibility can be used to minimize the fluctuation of (2) about its average value by optimizing its variance about that value. The minimum variance controllers are of this latter type.

In this rather brief discussion, it is notationally convenient to designate (2) simply as J, its average value by M, and the variance of J about M by V. Thus

$$E\{ J \} = M , \tag{4}$$

$$E\{ (J-M)^2 \} = V . \tag{5}$$

The minimum variance controller seeks to satisfy (4) and to minimize (5). Depending upon the information which is made available to the controller, various solutions are possible.

SOLUTION

Feedback solutions of this problem have been considered [1,2]. This paper, however, is addressed to the improvement of the open loop solution, which was first described in 1965-66 [3,4]. The solution proposed in [4] was adequate in principle, but quite tedious in practice. It was based upon obtaining the characteristic values of a coupled set of linear integro-differential equations. The characteristic values were then used to form an assumed solution with unknown coefficients; finally the coefficients in the solution were determined by substituting into the original equations.

The principal drawback of that method was the necessity of carrying out multiple hand calculations before the equations could be put in a form suitable for the computer. Although certain qualitative properties of the relationship between M and V were discovered [3], the above mentioned difficulties limited numerical studies to the first order case [4]. Even the first-order studies of [4], moreover, did not contain trajectory and control action information in graphic form.

As a result of recent work by Baggeroer, however, it is now possible to convert the integro-differential equations of [4] to coupled differential equations. This simple conversion places the problem solution in a form directly amenable to straightforward computer solution. An additional feature of the new approach is the rather easily obtained information on trajectories and control actions. Because of space limitations, only the precise form of the solution equations are stated here.

Following the work of Baggeroer [5], and after completing a derivation of the same type as in [4], it can be shown that the minimum variance, open loop controller solution is defined by the differential equations

$$\dot{z}(t) = Az(t) - \frac{1}{2\mu} BR^{-1}B'\rho(t) , \tag{6}$$

$$\dot{\rho}(t) = -A'\rho(t) - 2\mu Qz(t) - 8Qv(t) , \tag{7}$$

$$\dot{v}(t) = Av(t) + Sy(t) , \tag{8}$$

$$\dot{y}(t) = -A'y(t) - Qz(t) , \tag{9}$$

together with the boundary conditions

$$z(0) = x(0) , \tag{10}$$

$$\rho(t_f) = 2\mu Pz(t_f) + 8Pv(t_f) , \tag{11}$$

$$v(0) = 0 , \tag{12}$$

$$y(t_f) = Pz(t_f) , \tag{13}$$

and the control action relationship

$$u(t) = -\frac{1}{2\mu} R^{-1}B'\rho(t) . \tag{14}$$

The variable $z(t)$ is the mathematical expectation of $x(t)$; the variable $\rho(t)$ corresponds closely to the usual costate variable of optimal control theory, since it is the variable which enforces the differential equation constraint between $z(t)$ and $u(t)$. The variables $v(t)$ and $y(t)$ are introduced via the method of Baggeroer to reduce the integro-differential equation of [4] to the differential equations of (6) through (9). The parameter μ is positive and is the Lagrange-type variable used to enforce the constraint (4) on average value of J. In practice it turns out to be convenient to solve the

equations (6) through (14) for several values of μ, so that insights can be obtained into the possible tradeoffs between average value M and the minimum variance V about it. In general, large μ corresponds to small M, and small μ to large M. The values of M and V are given by

$$M = z'(t_f)Pz(t_f) + \int_0^{t_f}[z'(\sigma)Qz(\sigma) + \frac{1}{4\mu^2}\rho'(\sigma)BR^{-1}B'\rho(\sigma)]d\sigma + K_m \qquad (15)$$

$$V = 4[z'(t_f)Pv(t_f) + \int_0^{t_f}[z'(\sigma)Qv(\sigma)]d\sigma + K_v \ , \qquad (16)$$

where K_m and K_v are the constant mean and variance of a random variable not influenced in any way by the controller. K_m and K_v are not computed herein.

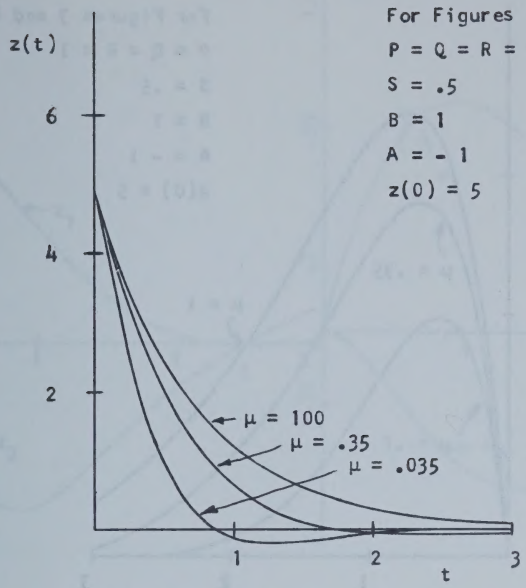
NUMERICAL RESULTS

Figures 1 through 10 give the curves of solution of (6) through (14) for representative systems of first, second, and fifth order, the last system being the well known example of Athans and Levine [6]. In perusing the curves, it is well to remember that z(t) is the average state trajectory and that $\rho(t)$ is its costate. The parameter μ, used here as a design quantity, is inversely related to M and directly related to V. The table below gives M and V values for the examples of Figures 1 through 10.

	Figures 1-4		Figure 5		Figures 6-7		Figure 8	
μ	M	V	M	V	M	V	M	V
.035	31.9	.54			66.2	4.0		
.1	23.4	1.04			45.7	5.2		
.35	18.0	2.04	159	35.2	33.2	7.5		
1.0	16.2	3.03	133	50.8	28.5	10.2		
10							1908	3748
100	15.6	4.53			25.8	18.0		

REFERENCES

1. M. K. Sain, "A Sufficiency Condition for Minimum Variance Control of Markov Processes," Proc. Fourth Allerton Conf. on Circuit and System Theory, pp. 593-599, October 1966.
2. S. R. Liberty and M. K. Sain, "Minimum Variance Feedback Controllers: Initial Studies of Solutions and Properties," Proc. Sixth Allerton Conf. on Circuit and System Theory, pp. 408-417, October 1968.
3. M. K. Sain, "Relative Costs of Mean and Variance Control for a Class of Linear, Noisy Systems," Proc. Third Allerton Conf. on Circuit and System Theory, pp. 121-129, October 1965.
4. M. K. Sain, "Control of Linear Systems According to the Minimal Variance Criterion--A New Approach to the Disturbance Problem," IEEE Trans. on Automatic Control, Vol. AC-11, pp. 118-121, January 1966.
5. A. B. Baggeroer, "A State Variable Approach to the Solution of Fredholm Integral Equations," IEEE Trans. on Information Theory, Vol. IT-15, pp. 557-570, September 1969.
6. W. S. Levine and M. Athans, "On the Optimal Error Regulation of a String of Moving Vehicles," IEEE Trans. on Automatic Control, Vol. AC-11, pp. 355-361, July 1966.



For Figures 1 and 2:
P = Q = R = 1
S = .5
B = 1
A = - 1
z(0) = 5

Figure 1

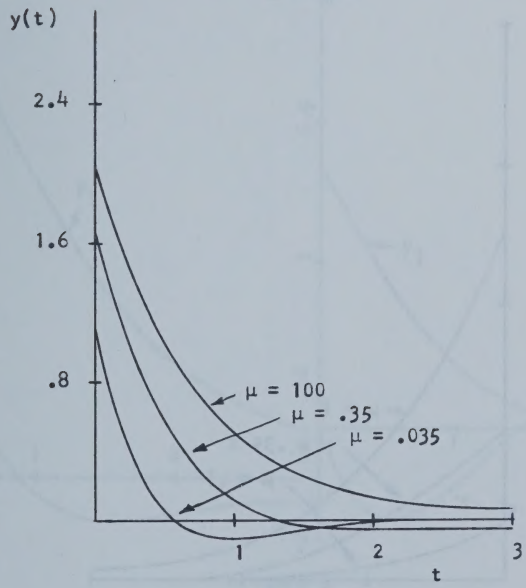


Figure 2

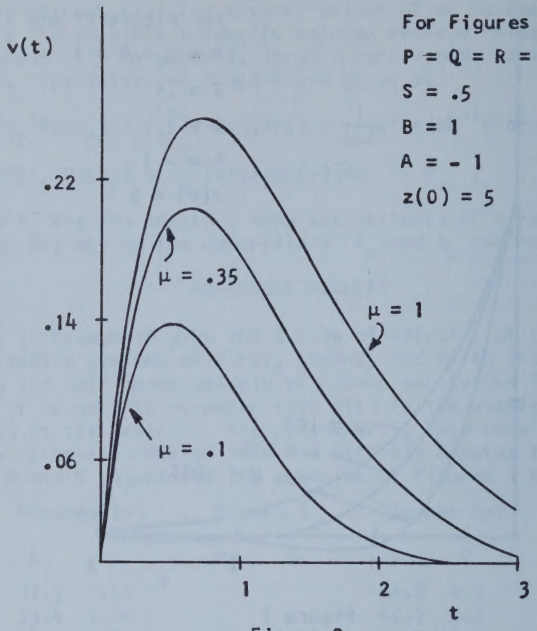


Figure 3

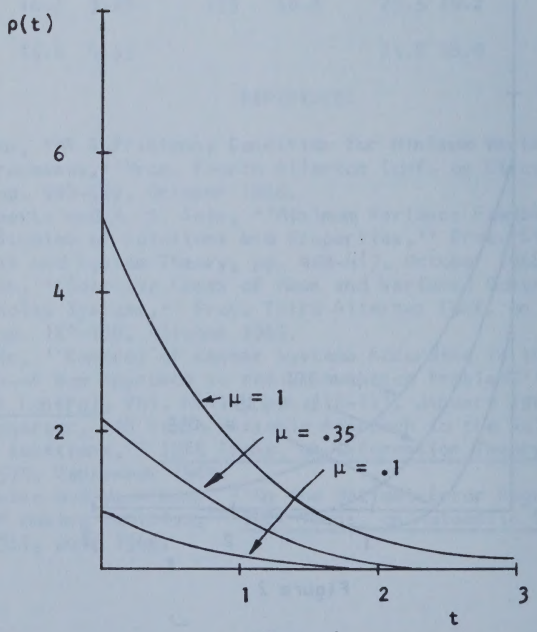


Figure 4

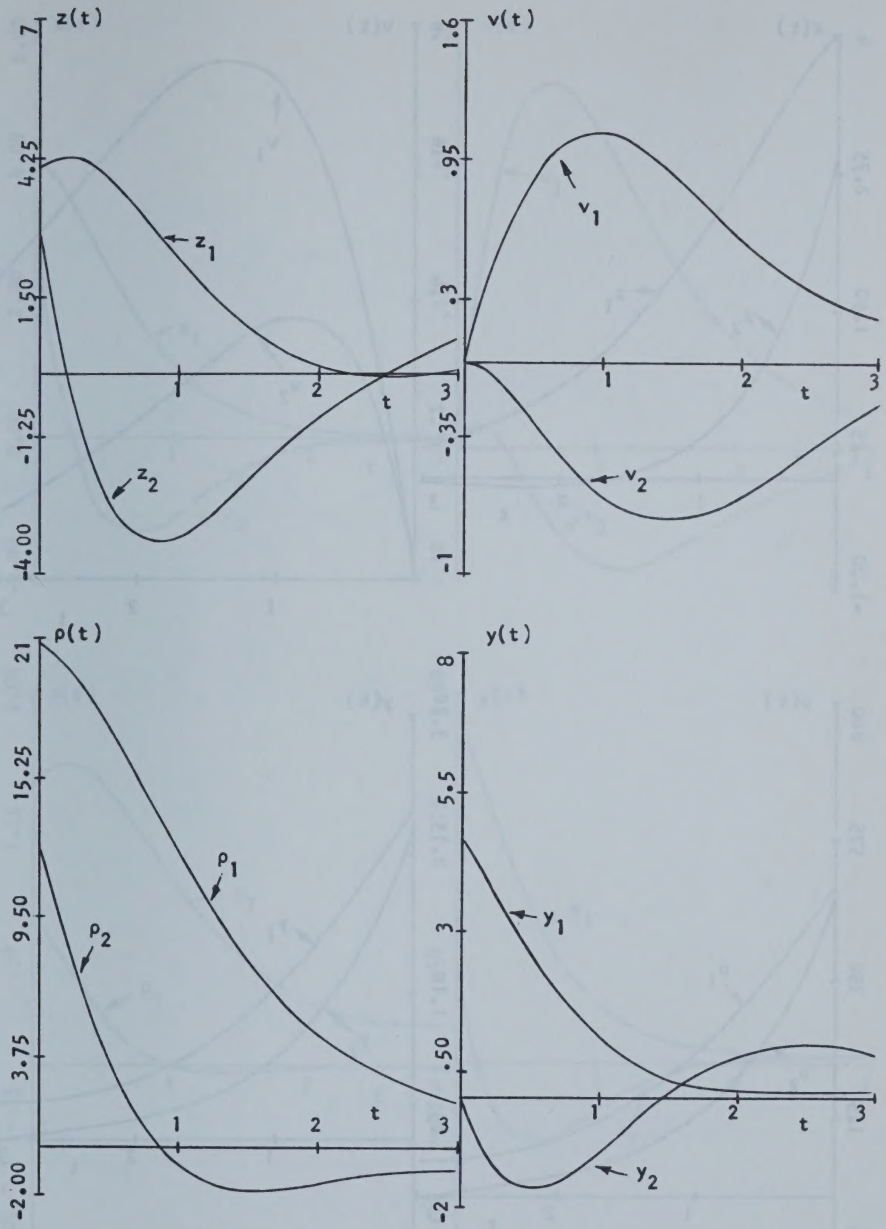


Figure 5

$$A = \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix} \quad B = \begin{vmatrix} 0 \\ 1 \end{vmatrix} \quad P = Q = R = I \quad S = .5I$$

$$\mu = .35$$

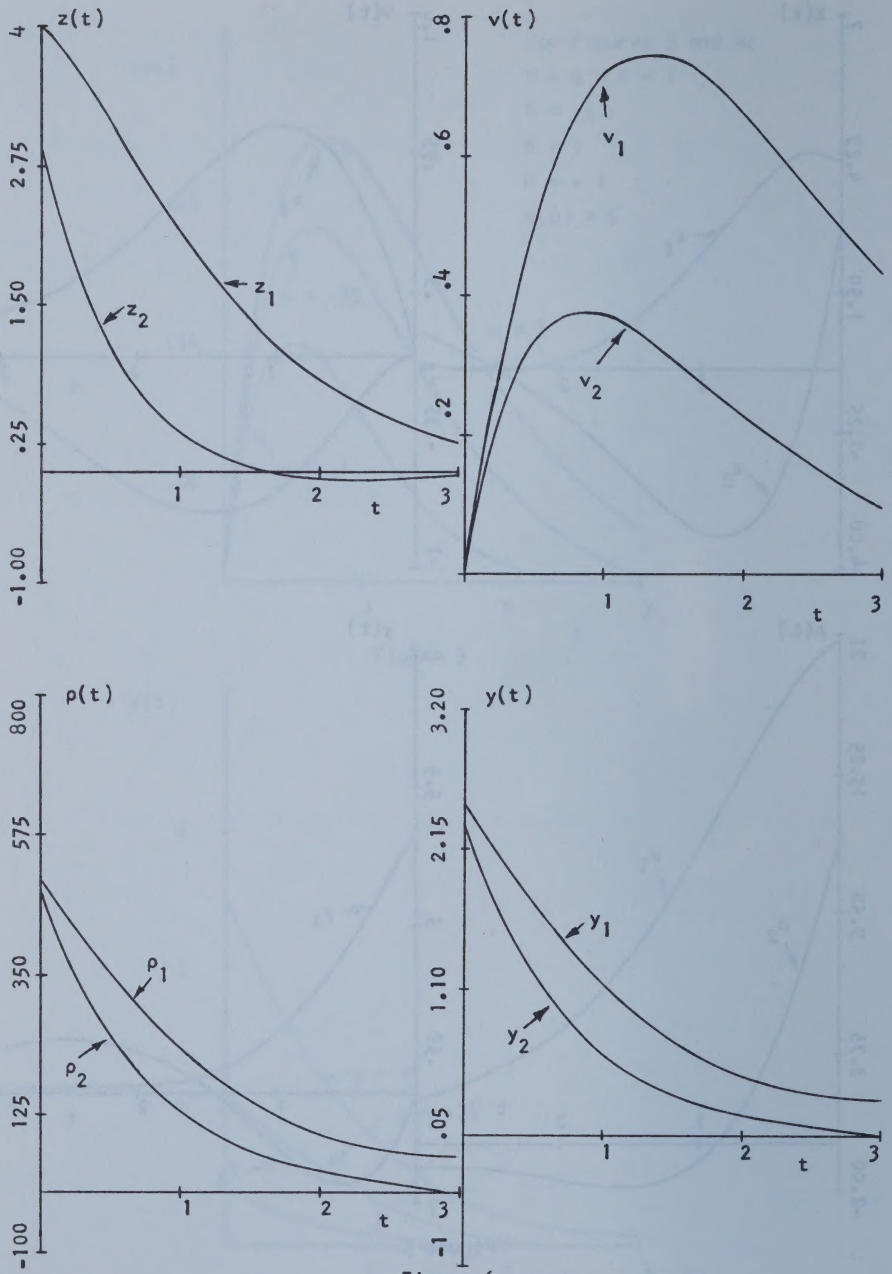


Figure 6

$$A = \begin{vmatrix} -1 & 1 \\ 0 & -1 \end{vmatrix} \quad B = \begin{vmatrix} 0 \\ 1 \end{vmatrix} \quad P = Q = R = I \quad S = .5I$$

$$\mu = 100$$

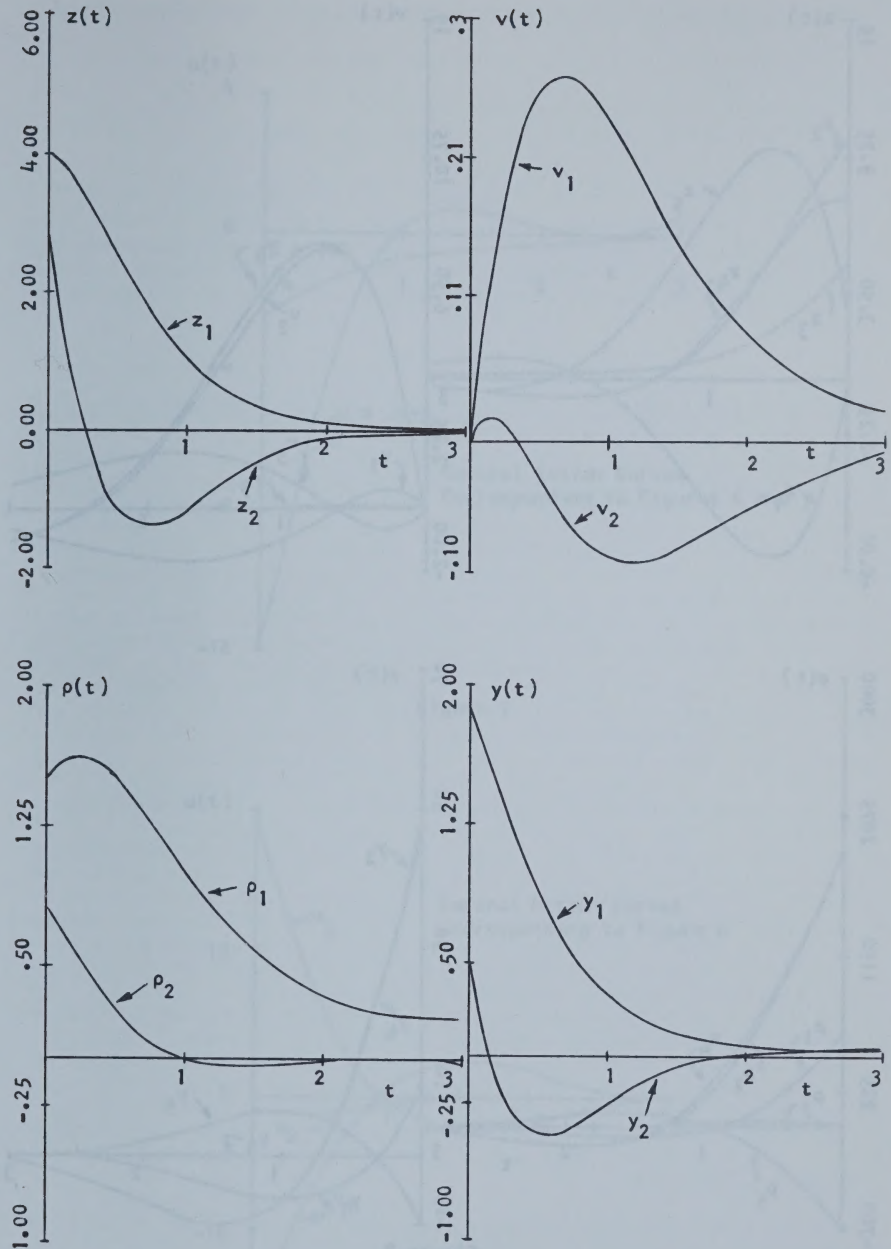


Figure 7

$$A = \begin{vmatrix} -1 & 1 \\ 0 & -1 \end{vmatrix} \qquad B = \begin{vmatrix} 0 \\ 1 \end{vmatrix} \qquad P = Q = R = I \qquad S = .5I$$

$$\mu = .035$$

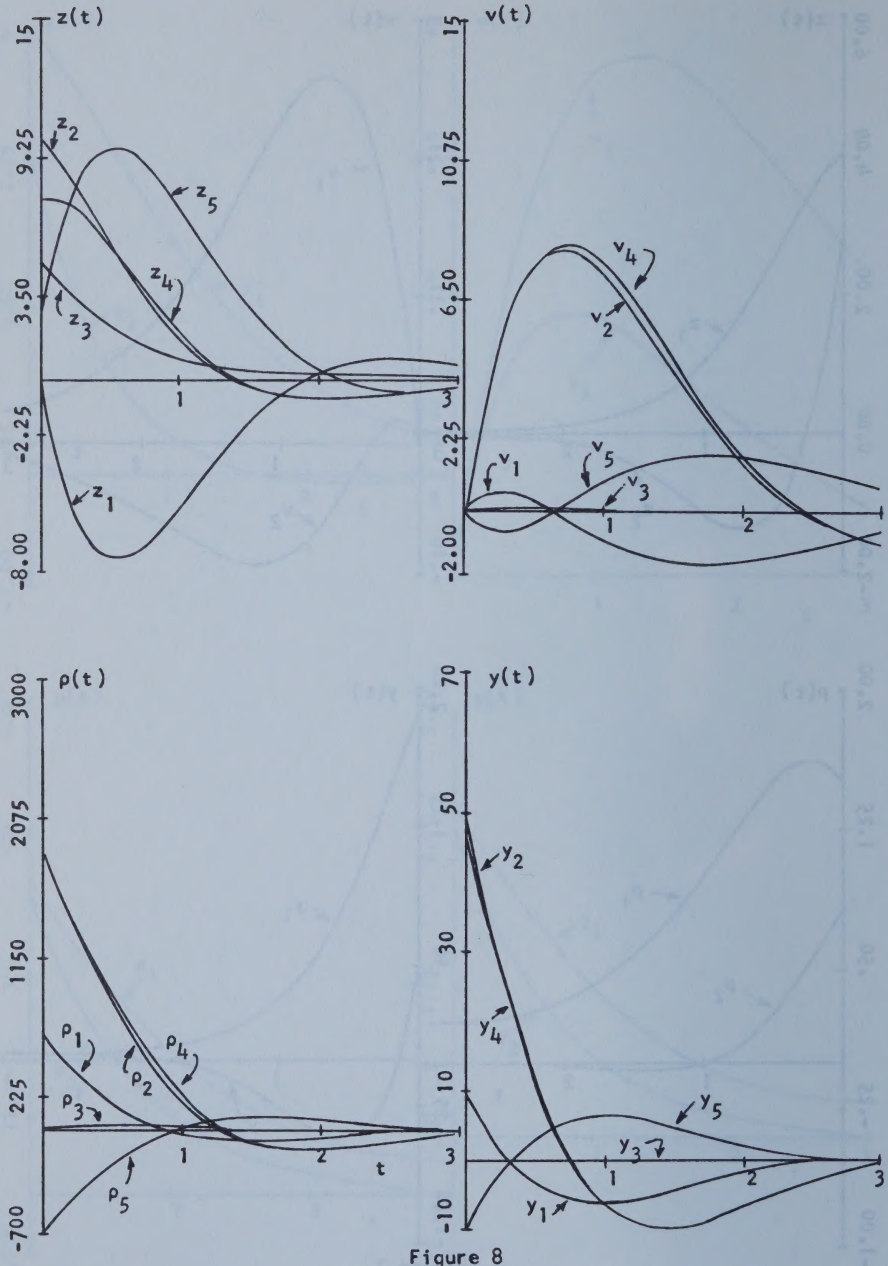
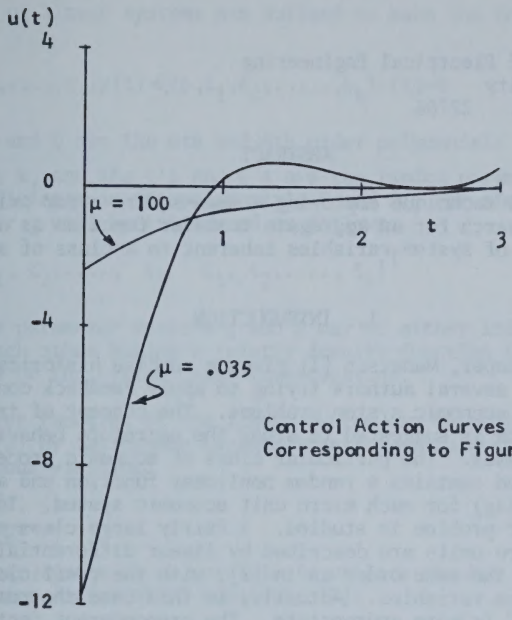


Figure 8

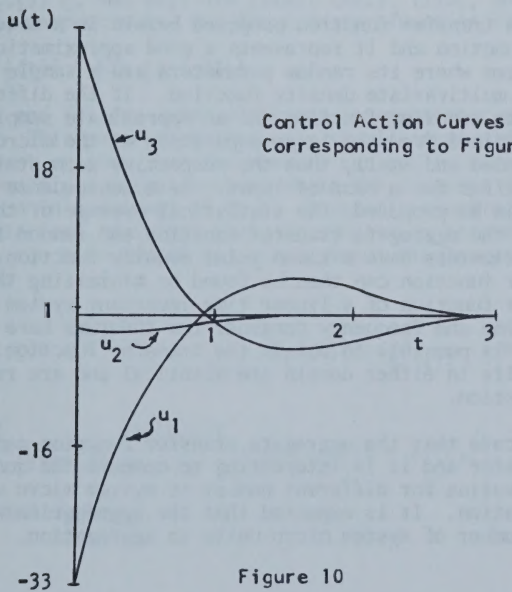
A, B, P, Q, R as in Athans-Levine reference; S = .5I

$\mu = 10$



Control Action Curves
Corresponding to Figures 6 and 7

Figure 9



Control Action Curves
Corresponding to Figure 8

Figure 10

AN ANALYTICAL STUDY OF THE AGGREGATE TRANSFER FUNCTION FOR LINEAR STOCHASTIC SYSTEMS

PAUL P. WANG
Department of Electrical Engineering
Duke University
Durham, N. C. 27706

ABSTRACT

A statistical technique employing a mean-square-error criterion is utilized herein to search for an aggregate transfer function as a means of reducing the number of system variables inherent to a class of stochastic linear systems.

I. INTRODUCTION

In a recent paper, Manetsch [1] gives a concise historical description of the efforts of several authors trying to apply feedback control theory to management and economic system problems. The concept of transfer function representation is suggested to study the aggregate behavior of a class of economic processes. The particular class of economic processes that Manetsch considered contains a random nonlinear function and a random pure delay (transport lag) for each micro unit economic system. In this paper a relatively similar problem is studied. A fairly large class of economic systems whose micro units are described by linear differential equations of the same form and the same order as in [2], with the coefficient parameters representing random variables. Actually, in this case the name "aggregate transfer function" is more appropriate. The preponderant fact is that many economic control systems such as inventory systems (factory warehouse problems) are inherently linear [2],[3], alternatively, nonlinear economic systems can be approximated by a linear model either by first order perturbation or linear approximation when the nonlinearity is not severe, [5],[11].

The aggregate transfer function proposed herein is a deterministic (function of t) function and it represents a good approximation to a micro unit economic system where its random parameters are a sample vector drawn from a particular multivariate density function. If the difference between the exact aggregate transfer function and an approximate sample transfer function, each obtained from the large population of the micro unit economic system, is bounded and small, then the respective zero state responses will definitely differ for a bounded input. As a consequence the following error criterion can be proposed: the statistical average of the square of the difference of the aggregate transfer function and random transfer function where its parameters have a known joint density function. The exact aggregate transfer function can then be found by minimizing this criterion. Since the transfer function of a linear time-invariant system can be expressed both in time and frequency domains, two theorems have been established so that it is possible to obtain the transfer function in either domain. The results in either domain are identical and are related by the Laplace transformation.

Results indicate that the aggregate transfer function can be viewed as an unbiased estimator and it is interesting to compare the quality of the aggregate approximation for different number of system micro units in the process of aggregation. It is expected that the aggregation will improve with increasing number of system micro units in aggregation.

II. AGGREGATE TRANSFER FUNCTION FOR A CLASS OF LINEAR STOCHASTIC SYSTEMS

The class of linear systems are defined to have the following general form:

$$P(p,\zeta_1,\zeta_2,\ldots,\zeta_j)y(t)-Q(p,\xi_1,\xi_2,\ldots,\xi_k)u(t)=0 \tag{1}$$

where $p \triangleq \frac{d}{dt}$, P and Q are the nth and mth order polymonials in p, respectively, with $n \geq m$, and the ζ's and ξ's are the random parameters which appear as coefficients in the polynomials P and Q. For simplicity, the parameter vector $\underline{\zeta}$ will be introduced as follows:

$$\underline{\zeta}^T=[\underline{\zeta}^T \quad \underline{\xi}^T]=[\zeta_1, \zeta_2,\ldots, \zeta_j \quad \xi_1, \xi_2,\ldots, \xi_k]^T$$

The random parameter vectors $\underline{\zeta}$ and $\underline{\xi}$ may be either independent or correlated with each other having a jointly density function $f_{\underline{\zeta}}(\underline{\zeta})$ dependent upon the nature of the system under investigation.

The transfer function of the linear system can be obtained immediately from equation (1) by taking the Laplace transformation and assuming zero initial conditions yielding:

$$H(s) = \frac{Q(s;\ \underline{\xi})}{P(s;\ \underline{\zeta})} \tag{2}$$

For a linear system it is possible to utilize another form for the first order vector differential equation, [8]:

$$\dot{\underline{z}}(t) = \underline{A}(\underline{\zeta})\ \underline{z}(t) + \underline{b}(\underline{\zeta},\underline{\xi})u(t) \tag{3a}$$

$$y(t) = \underline{c}\underline{z}(t) + d(\underline{\xi})u(t) \tag{3b}$$

where $\underline{A}(\underline{\zeta})$, $\underline{b}(\underline{\zeta},\underline{\xi})$, $\underline{c}$, and $d(\underline{\xi})$ are (nxn), (nx1), (1xn), and (1x1) dimensioned matrices, respectively, and the transfer function is, [4],

$$H(s)=\underline{c}[s\underline{I}-\underline{A}(\underline{\zeta})]^{-1}\ \underline{b}(\underline{\zeta},\underline{\xi}) + d(\underline{\xi}) \tag{4}$$

Even though (2) and (4) are quite different in appearance they must be, and are, identical. This paper considers only single input-single output linear time invariant systems, but the results can easily be extended to multiple input-multiple output systems where the transfer function will be a matrix function in s instead of in t. Generally speaking, the poles of (2) can be single real, multiple real, or complex in nature. The transfer function in the time domain for single input-single output systems can be expresses as:

$$h(t) = \sum_{i=1}^{s_1} K_i(\underline{\zeta},\underline{\xi})e^{\lambda_i(\underline{\zeta},\underline{\xi})t} + \sum_{i=1}^{s_2} N_i(\underline{\zeta},\underline{\xi})\ \frac{t^{i-1}}{(i-1)!}\ e^{\lambda_{s_2}(\underline{\zeta},\underline{\xi})t}$$

$$+ \sum_{i=1}^{n-s_1-s_2} T_i(\underline{\zeta},\underline{\xi})e^{\Omega_i(\underline{\zeta},\underline{\xi})t} \cos[\omega_i(\underline{\zeta},\underline{\xi})t + \phi_i(\underline{\zeta},\underline{\xi})] \tag{5}$$

where $s_1 \varepsilon [o,n]$ and $s_2 \varepsilon [o,n]$.

Intuitive arguments will now be presented in order to demonstrate obtaining the concept of aggregating the transfer function. Since the system is linear, a sample zero-state output of a sample system $y_{\underline{\zeta}}(t)$ excited by input $u(t)$ can be found through the convolution integration

$$y_{\underline{\zeta}}(t) = \int_0^t u(t-\tau)h(\tau;\underline{\zeta},\underline{\xi})d\tau \tag{6}$$

It is possible to find a deterministic aggregate transfer function $\theta(t)$, such that

$$\max_{t\in[0,\infty]} |\theta(t) - h(t;\underline{\zeta},\underline{\xi})| \leq M \qquad \forall t \tag{7}$$

where M is a finite real number, and (7) is valid for the whole class of random transfer functions of the large scale stochastic system. In addition, the zero state response for the system having transfer function $\theta(t)$ excited by the same input $u(t)$ is;

$$y_{\theta}(t) = \int_0^t u(t-\tau)\,\theta(\tau)d\tau. \tag{8}$$

The difference, which is the quantity of interest, can be established immediately as follows:

$$y_{\underline{\zeta}}(t) - y_{\theta}(t) = \int_0^t u(t-\tau)[h(\tau;\underline{\zeta},\underline{\xi})-\theta(\tau)]d\tau \tag{9}$$

$$\leq \int_0^t u(t-\tau) \max_{t\in[0,\infty]} |\theta(\tau)-h(\tau;\underline{\zeta},\underline{\xi})|\, d\tau$$

$$= M \int_0^t u(t-\tau)d\tau \tag{10}$$

This expression is bounded since it is assumed that the input $u(t)$ in reality is greater than zero for all time and is a bounded piecewise continuous function. Optimally, if $\theta(t)$ can be chosen such that M is minimum, then the existence of the difference $y_{\underline{\zeta}}(t) - y_{\theta}(t)$ for a specific sample can also be guaranteed in statistical sense. Therefore it is desirable to evaluate $\theta(t)$ under a preselected criterion such that it is optimal in the statistical sense. The criterion proposed herein is the well known mean-square-error criterion where the quantity to be minimized is given by:

$$[\theta(t)-h(t;\,\zeta_1,\zeta_2,\ldots,\zeta_j,\xi_1,\xi_2,\ldots,\xi_k)]^2 \text{ for all } t\in[0,\infty)$$

where the aggregate transfer function $\theta(t)$ is the desired function. Since the problem entails a class of transfer functions and it is only meaningful to talk about some statistical averaging criterion, it is therefore desirable to minimize the averaged M.S.E..

$$\underset{\theta(t),\ t\in(0,\infty)}{\text{Min}}\ g[\theta(t)] = \text{Min}\ E\,[\theta(t) - h(t;\underline{\zeta},\underline{\xi})]^2$$

$$= \min \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\theta(t)-h(t;\zeta_1,\zeta_2,\ldots,\zeta_j,\xi_1,\xi_2,\ldots,\xi_k)]^2 \quad (j+k \text{ integrals})$$

$$f_{\zeta_1,\zeta_2,\ldots,\zeta_j,\xi_1,\xi_2,\ldots,\xi_k}(\zeta_1,\zeta_2,\ldots,\zeta_j,\xi_1,\xi_2,\ldots,\xi_k) \cdot$$

$$d\zeta_1 d\zeta_2 \ldots d\zeta_j d\xi_1 \ldots d\xi_k. \qquad (11)$$

For simplicity, (11) may be rewritten as:

$$\underset{\theta(t),t\in(0,\infty)}{\text{Min}}\ g\,[\theta(t)] = \min \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\theta(t)-h(t;\underline{\zeta},\underline{\xi})]^2 f_{\underline{C}}(\underline{c})d\underline{c} \qquad (12)$$

where $f_{\underline{C}}(\underline{c})$ is the joint density function for the parameter vector $\underline{c}$. If $\underline{\zeta}$ and $\underline{\xi}$ are statistically independent, then

$$f_{\underline{C}}(\underline{c}) = f_{\underline{\zeta}}(\underline{\zeta})\ f_{\underline{\xi}}(\underline{\xi}) \qquad (13)$$

Under the above specific criterion, the minimization process yields the solution for $\theta(t)$ which is found to be the averaged time function of the stochastic function $h(t;\underline{\zeta},\underline{\xi})$ as summarized in the following theorem:

Theorem 1: Given the criterion of (11), the time-domain aggregate transfer function $\theta(t)$ is the statistical average of $h(t;\underline{\zeta},\underline{\xi})$.

[Proof] Taking the derivative of (12) with respect to $\theta(t)$ yields

$$\frac{\partial g[\theta(t)]}{\partial \theta(t)} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 2[\theta(t)-h(t;\underline{\zeta},\underline{\xi})]\ f_{\underline{C}}(\underline{c})d\underline{c} = 0, \quad \forall t \qquad (14)$$

or

$$\theta(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(t;\underline{\zeta},\underline{\xi}) f_{\underline{\zeta},\underline{\xi}}(\underline{\zeta},\underline{\xi})d\underline{\zeta}\ d\underline{\xi} \qquad (15)$$

[Alternative Proof] If $h(t;\underline{\zeta},\underline{\xi})$ is regarded as another random variable $\bar{x}(t)$, (11) can be written as

$$\min_{\theta(t)} g[\theta(t)] = \min_{\theta(t)} E[\theta(t) - \bar{x}(t)]^2, \quad \forall t$$

$$= \min \{ \text{Var.}[\bar{x}(t)] + [E(\bar{x}(t)) - \theta(t)]^2 \} \qquad (16)$$

or,

$$\min_{\theta(t)} g[\theta(t)] = \text{Var.}\ [\bar{x}(t)] \tag{17}$$

since,

$$\theta(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(t;\underline{\zeta},\underline{\xi}) f_{\underline{\zeta},\underline{\xi}}(\underline{\zeta},\underline{\xi}) d_{\underline{\zeta}},d_{\underline{\xi}} = E[\bar{x}(t)] \tag{18}$$

$j+k$

Taking the Laplace transform of (18), the aggregate transfer function is obtained in the frequency domain. It will, however, be proven that frequency-domain aggregate transfer function $\Theta(s)$ can be obtained directly from frequency-domain formulation, i.e., $\Theta(s)$ is the statistical average of the random function $H(s;\underline{\zeta},\underline{\xi})$, where

$$H(s;\underline{\zeta},\underline{\xi}) = L[h(t;\underline{\zeta},\underline{\xi})].$$

Theorem II: Given the criterion of (11) in frequency domain, the aggregate transfer function $\Theta(s)$ can be found by taking statistical average of $H(s;\underline{\zeta},\underline{\xi})$.

[Proof] The criterion of (11) expressed in the frequency domain yields

$$\min_{\Theta(s)} G[\Theta(s)] = \min_{\Theta(s)} E[\Theta(s) - H(s;\underline{\zeta},\underline{\xi})]^2 \tag{19}$$

The quantity inside the parenthesis of (19) is a complex variable. The true meaning of the minimum of a complex variable is not quite clear at this point. It is maintained that if $\Theta(s) = \alpha + j\beta$ and the stochastic transfer function to be aggregated is viewed as another random variable $\bar{z}(s) = H(s;\underline{\zeta},\underline{\xi}) = \bar{z}_1 + j\bar{z}_2$, then $\Theta(s) = \alpha + j\beta = E[\bar{z}_1 + j\,\bar{z}_2]$ (20)

or

$$\Theta(s) = E[\bar{z}(s)] = E[H(s;\underline{\zeta},\underline{\xi})] \tag{21}$$

To prove equations (20) or (21) note that

$$\begin{aligned} E[\Theta(s) - \bar{z}(s)]^2 &= E[\bar{z}_1 + j\,\bar{z}_2 - \alpha - j\beta]^2 \\ &= \text{Var.}\ [\bar{z}_1] - \text{Var.}\ [\bar{z}_2] + 2j\ \text{Cov.}[\bar{z}_1,\bar{z}_2] \\ &\quad + (E[\bar{z}_1] + j\ E[\bar{z}_2] - \alpha - j\beta)^2 \end{aligned} \tag{22}$$

The first three terms of (22) contains the variances and covariance of random variables, therefore, the criterion (22) is minimized if and only if

$$E[\bar{z}_1] + j\ E[\bar{z}_2] - \alpha - j\beta = 0 \tag{23}$$

or

$$\theta(s) = E[\bar{z}(s)] = E[H(s;\underline{\zeta},\underline{\xi})]$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(s;\underline{\zeta},\underline{\xi}) f_{\underline{\zeta},\underline{\xi}}(\underline{\zeta},\underline{\xi}) d\underline{\zeta}\, d\underline{\xi} \qquad (24)$$

$j+k$

As previously mentioned (24) could have been justified by taking the Laplace transformation of (15). However, since the complex random variables are involved, it is more satisfactory to justify the solution by the more rigorous proof of Theorem II.

Corollary I: The variance of the time-domain stochastic transfer function can be evaluated by utilizing either Theorem I or II.

[Proof] Designate the variance, which is a function of time, as $\sigma^2(t)$, then:

$$\sigma^2(t) = E\{[h(t;\underline{\zeta},\underline{\xi}) - E(h(t;\underline{\zeta},\underline{\xi}))]^2\}$$

$$= E[h^2(t;\underline{\zeta},\underline{\xi})] - \{E[h(t;\underline{\zeta},\underline{\xi})]\}^2 \qquad (25)$$

Utilizing Theorem I, it is concluded that:

$$\sigma^2(t) = E[h^2(t;\underline{\zeta},\underline{\xi})] - \theta^2(t)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h^2(t;\underline{\zeta},\underline{\xi}) f_{\underline{\zeta},\underline{\xi}}(\underline{\zeta},\underline{\xi}) d\underline{\zeta}\, d\underline{\xi} - \theta^2(t) \qquad (26)$$

$j+k$

Corollary II: If $H_V(s;\underline{\zeta},\underline{\xi}) = L[h^2(t;\underline{\zeta},\underline{\xi})]$ and $\Theta_V(s) = L[\theta^2(t)]$, then the variance of the frequency-domain stochastic transfer function is:

$$\Gamma_V(s) = E[H_V(s;\underline{\zeta},\underline{\xi})] - \Theta_V(s)$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H_V(s;\underline{\zeta},\underline{\xi}) f_{\underline{\zeta},\underline{\xi}}(\underline{\zeta},\underline{\xi}) d\underline{\zeta}\, d\underline{\xi} - \Theta_V(s) \qquad (27)$$

$j+k$

The proof of Corollary II is straightforward since it is known that $\Gamma_V(s) = L[\sigma^2(t)]$. It is worthwhile to note that $H_V(s;\underline{\zeta},\underline{\xi})$ is not equal to the square of $H(s;\underline{\zeta},\underline{\xi})$, unlike Theorems I and II, since the procedure is no longer a linear operation.

III. APPLICATIONS

Referring to (1), it should be noted that if u(t) is a delta function, $\delta(t)$, and statistical averaging has been applied, then the aggregate transfer function can be interpreted as a function of time, $\theta(t) = y(t)$. Once $\theta(t)$ is obtained, then the general deterministic time function u(t) can be used as an input to a large class of systems. Another important problem is the treatment of zero-input responses. In the Gauss-Markov sequence model, the n-dimensional initial state vector is assumed to be a gaussian

- random vector with known mean and covariance matrix, [7]. In general, the zero-input response has significant physical interpretation [2],[3], and thus the existence of the initial state vector certainly can not be ignored [10]. Considering the stochastic transition matrix of (3) to be $\Phi(t;t_o;\underline{\zeta},\underline{\xi})$, then the zero input response if [8]:

$$y(t) = [1\ 0\ \ldots\ 0]\underline{z}(t) = [1\ 0\ \ldots\ 0]\ \underline{\Phi}(t,t_o;\underline{\zeta},\underline{\xi})\underline{z}(0)$$

$$= [1\ 0\ \ldots\ 0]\ \underline{\Phi}(t,t_o;\underline{\zeta},\underline{\xi})\ \underline{\gamma}(0) \qquad (28)$$

If the random vector $y(0)$ is independent of the random system parameter vector $\underline{\zeta}$, then:

Proposition I: If the random initial state vector and the random system parameters are stochastically independent, then the aggregation for the zero input response of a linear system can be obtained by aggregating the initial state vector and stochastic transfer function separately.

[Proof] It is trivial to see in (28) that $y(0)$ is independent of $\underline{\Phi}(t,t_o;\underline{\zeta},\underline{\xi})$.

IV. EXAMPLES

In this section two examples are presented which are illustrative of the theory developed in this paper.

[Example I] Given a very simple stochastic transfer function containing a random parameter γ:

$$h(t,\gamma) = e^{-\gamma t} = L^{-1}\left[\frac{1}{s+\gamma}\right]. \qquad (29)$$

The time-domain aggregate transfer function can be evaluated using (15), assuming $\gamma\varepsilon[0,\infty)$,

$$\theta_1(t) = \int_0^\infty e^{-\gamma t}\ f(\gamma)d\gamma = L[f(\gamma)] \qquad (30)$$

Equivalently, the frequency-domain aggregate transfer function can be evaluated using (24),

$$\Theta_1(s) = \int_0^\infty \left(\frac{1}{s+\gamma}\right)\ f(\gamma)\ d\gamma \qquad (31)$$

In order to prove (30) and (31) yield the same result, assume $f(\gamma) = \beta e^{-\beta\gamma}$ then (30) yields:

$$\theta_1(t) = L[\beta e^{-\beta\gamma}] = \frac{\beta}{t+\beta}\ ;\ \beta > 0 \qquad (32)$$

and therefore

$$\Theta_1(s) = L[\theta_1(t)] = L\left[\frac{\beta}{t+\beta}\right] = \beta e^{\beta s}\ E_1(\beta s) \qquad (33)$$

where $E_1(\beta s)$ is the so called exponential integral function [9],

$$E_1(\beta s) = \int_{\beta s}^{\infty} \frac{e^{-t}}{t}\, dt$$

Performing the actual integration in (31) yields exactly the same result as in (33). Thus the equivalence of Theorems I and II is proven valid.

Now imagine that a very large number of linear systems described by the above aggregate transfer function $\theta_1(s)$ is excited by an aggregate or equivalent input $\frac{A}{s}$, then the response can be computed:

$$y(t) = A\, u(t) * \frac{\beta}{t+\beta} = A\beta \ln \frac{t+\beta}{\beta} \tag{34}$$

where * represents the convolution integration.

[Example II] The stochastic transfer function for a simple first order linear system is $\underline{x}\, e^{-\underline{\gamma} t}$. Assume $\underline{x}$ is expontentially distributed with mean $\frac{1}{\alpha}$, i.e., $f_1(\underline{x}) = \alpha e^{-\alpha \underline{x}}$ and $\underline{\gamma}$ is uniformly distributed with magnitude $\frac{1}{\beta}$ where $\underline{\gamma} \in [0, \beta]$. For a specific case, $\alpha = 5$ and $\beta = 2$, the time-domain aggregate transfer function is found to be:
$\theta_2(t) = \frac{1-e^{-2t}}{10t}$, and using (26) the variance of the stochastic transfer function is:

$$\sigma_2^2(t) = \int_0^{\infty} \int_0^{\infty} (\underline{x}\, e^{-\underline{\gamma} t})^2 f_1(\underline{x})\, f_2(\underline{\gamma})\, d\underline{x}\, d\underline{\gamma} - \theta_2^2(t)$$

$$= \frac{1-e^{-4t}}{50\, t} - \frac{(1-e^{-2t})^2}{100\, t^2} \tag{35}$$

The variance $\sigma_2^2(t)$ is plotted in Figure 1 and it is seen that its value decreases rapidly as t becomes large. Figure 2 shows ten approximate aggregate transfer functions, which were obtained utilizing a random number generator, together with the accurate (theoretical) aggregate transfer function $\theta_2(t)$ for the purpose of comparison. Each approximated aggregate transfer function is obtained by averaging 25 independent stochastic transfer functions, this is the realistic situation such as would be encountered in a typical factory warehouse problem. The problem of approximation based upon only finite samples being available requires extensive treatment from the point of view of random sampling distribution theory. What is depicted here is a typical experiment using digital computer simulation.

V. CONCLUSIONS

An analytical study of the so defined aggregate transfer function is presented in this paper. The justification for the suggestion of solving a large class of linear stochastic system problems via aggregating its stochastic transfer functions using the expected MSE criterion is also discussed. Two theorems are derived which enable the aggregate transfer function either to be obtained from time-domain or frequency-domain approaches. The theoretical development of this paper is inspired by the

real need for a statistical aggregation technique for solving large scale economic systems [1] and operation research models [2],[3], as well as the group behavior of systems of different types [5]. Manetsch's scheme of aggregation e^{-st}, really can be considered as a special case of aggregation in the frequency domain as given in Theorem II and it can now be recognized that it actually originated from the mean-square-error criterion

Basically it is possible to classify general abstract systems in terms of whether the system is either linear or nonlinear in character. Herein is presented an investigation of linear stochastic systems. Within the framework of criterions and assumptions, the linear stochastic problem can be solved satisfactorily in general form. Nonlinear stochastic problems are not amenable to the aforementioned techniques due to the possible peculiar behavior of inherently nonlinear systems.

References

[1] Manetsch, Thomas J. "Transfer Function Representation of the Aggregate Behavior of a Class of Economic Processes." IEEE Transactions on Automatic Control. Vol. AC-11, Nov. 4, October, 1966. pp 693-8.

[2] Wang, P.P. "The Use of Modern Control Concepts in Inventory Theory, Part I - Formulation and Characteristics," to be presented at the 36th National Meeting of Operations Research Society Society of America at Miami, Florida, Nov. 1969.

[3] Wang, P.P. "The Use of Modern Control Concepts in Inventory Theory, Part II - Design and Optimization," to be presented at the 36th National Meeting of Operations Research Society Society of America, at Miami, Florida Nov. 1969.

[4] Hsu, J.C. and Meyer, A.U. Modern Control Principles and Applications McGraw-Hill Book Company, New York, 1968, pp. 67-68.

[5] Cosgniff, R.L. 'Solution of Statistical Problems by Automatic Control Techniques," 1957 IRE Conv. Rec., pt. 4, pp. 57-61.

[6] Feller, William, An Introduction to Probability Theory and its Applications, Volume 1, John Wiley & Sons, Inc., Third Edition, 1968.

[7] Meditch, J.S. Stochastic Optimal Linear Estimation and Control, McGraw-Hill Book Company, 1969, Chapter IV.

[8] Athans, M. and Falb, P.L. Optimal Control McGraw-Hill Book Company, 1966, Chapter 4.

[9] Abramowitz. Milton and Stegun, Irene A. Edited, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, U.S. Government Printing Office, P. 1029-1080.

[10] Liebelt, P.B., An Introduction to Optimal Estimation, Addison-Wesley Publishing Company, Inc., Reading, Mass., 1967 Chapter IV.

[11] Morrison, Norman, Introduction to Sequential Smoothing and Prediction. McGraw-Hill Book Company, New York, 1969, pp 105-116.

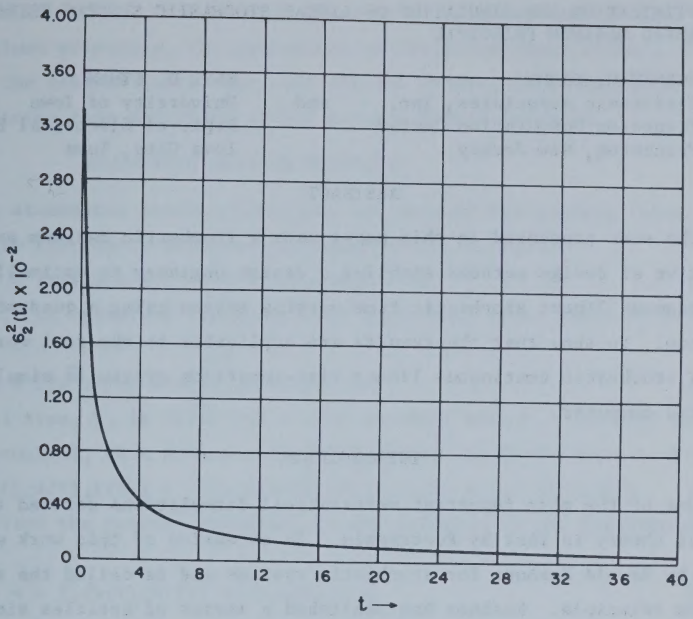


Figure 1. Variance of stochastic transfer function.

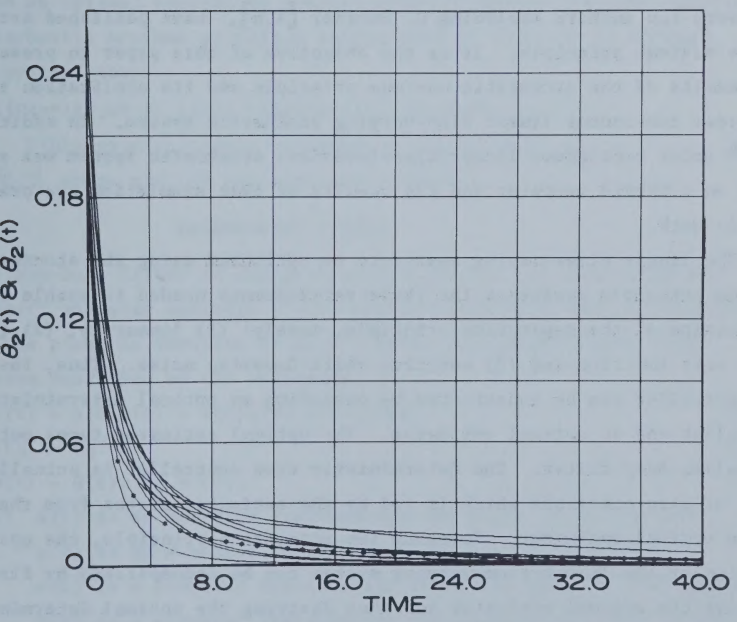


Figure 2. Theoretical $\theta_2(t)$ and approximate $\hat{\theta}_2(t)$ for sample size n=25.

THE OPTIMIZATION AND SIMULATION OF LINEAR STOCHASTIC SYSTEMS USING A STOCHASTIC MAXIMUM PRINCIPLE

NORMAN C. FROMM
Electronic Associates, Inc.
Princeton Computation Center
Princeton, New Jersey

and

EARL D. EYMAN
University of Iowa
Dept. of Electrical Engineering
Iowa City, Iowa

ABSTRACT

The work presented in this paper uses a stochastic maximum principle to arrive at design methods enabling a design engineer to optimally control a continuous linear stochastic time varying system using a quadratic cost function. To show that the results are applicable to the real world, an actual stochastic continuous linear time-invariant system is simulated on a hybrid computer.

INTRODUCTION

One of the most important mathematical formulations derived in optimal control theory is that by Pontryagin. An extension of this work was formulated by Harold Kushner for stochastic systems and is called the stochastic maximum principle. Kushner has published a series of articles since 1965 on the calculus [1], the existence [2], the proof of [3], and the sufficient conditions [4] for the stochastic maximum principle. At the present time very few authors excluding D. Sworder [5,6], have published articles on the maximum principle. It is the objective of this paper to present the results of the stochastic maximum principle and its application to an n^{th} order continuous linear time-varying stochastic system. In addition, an n^{th} order continuous linear time-invariant stochastic system was simulated on a hybrid computer and the results of that simulation are presented in this work.

The linear time-varying system to be optimized using the stochastic maximum principle possesses the three requirements needed to enable the application of the separation principle, namely: (1) linearity, (2) quadratic cost function and (3) additive white Gaussian noise. Thus, the optimal controller can be constructed by cascading an optimal deterministic controller and an optimal estimator. The optimal estimator turns out to be the Kalman-Bucy filter. The deterministic type controller is actually an array of gain constants which is fed by the estimated states from the output of the optimal estimator. Invoking the separation principle, the optimization of the linear time-varying system can be accomplished by first deriving the optimal estimator and then deriving the optimal deterministic controller.

This paper is divided into four parts, namely: (1) statement of the stochastic maximum principle, (2) derivation of the Kalman-Bucy filter, (3) derivation of the optimal controller, and (4) the results of a simulation of a continuous linear time-invariant system using a hybrid computer.

STOCHASTIC MAXIMUM PRINCIPLE

Kushner's stochastic maximum principle was derived for systems represented by an Ito stochastic differential equation of the form

$$\dot{x}(t) = f(x(t),u(t)) + g(t)w(t), x(t_o) = x_o. \tag{1}$$

This system will be optimized using the cost functional whose general form is

$$J = E\{c^T x(T)\} + E\{\int_o^T f_o(x(t),u(t))dt\} \tag{2}$$

where the final time, T, is fixed and c is a constant vector. Let u(t) be an admissible control, then define a function called the Hamiltonian to be

$$H(x(t),u(t),\lambda(t)) = \lambda^T(t)f(x(t),u(t)),u(t)) + f_o(x(t),u(t)) \tag{3}$$

where $\lambda(t)$, called the costate variable, is the solution of the differential equation

$$\dot{\lambda}(t) = - f_x^T(x(t),u(t))\ \lambda(t), \lambda(T) = c. \tag{4}$$

It can be shown that $\lambda(t)$ also satisfies the following equation

$$\lambda(t) = -\partial H(x(t),u(t),\lambda(t))/\partial x(t). \tag{5}$$

Let u*(t) be an optimal admissible control which minimizes the cost functional. The stochastic maximum principle states that if u*(t) is an optimal admissible control then

$$E\{H(x^*(t),u^*(t),\lambda(t)) \mid \text{information available at } t\} \leq E\{H(x(t),u(t),\lambda(t)) \mid \text{information available at } t\} \tag{6}$$

for all $t \in [0,T]$ where u(t) is any admissible control.

KALMAN-BUCY FILTER

The Kalman-Bucy filter is designed to extract from an observable process the "best" unbiased estimate of the state variables of the system being observed. The plant in question is an n^{th} order continuous linear time-varying system described by the relations

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + G(t)w(t) \tag{7}$$

$$x(t_o) = x_o \tag{8}$$

$$y(t) = H(t)x(t) + v(t) \tag{9}$$

where a) x(t) is an n-vector representing the state of the system,
u(t) is an m-vector representing the input of the system,
w(t) is a p-vector representing the noise input of the system,
v(t) is a r-vector representing the measurement noise of the system,

$A(t)$ is an nxn system matrix,

$B(t)$ is an nxm system input gain matrix,

$G(t)$ is an nxp system noise gain matrix, and

$H(t)$ is an rxn system output matrix.

where b) the system is completely controllable.

where c) the system noise is an additive white Gaussian noise process with zero mean

$$E\{w(t)\} = 0 \text{ for all } t \geq t_o \tag{10}$$

and a covariance given by

$$\text{cov}\left[w(t);\ w(\tau)\right] = Q(t)\ \delta(t - \tau) \tag{11}$$

where $Q(t)$ is a pxp symmetric positive semidefinite matrix.

where d) the measurement noise is an additive white Gaussian noise process with zero mean

$$E\{v(t)\} = 0 \text{ for all } t \geq t_o \tag{12}$$

and a covariance given by

$$\text{cov}\left[(v(t);\ v(\tau))\right] = N(t)\delta(t - \tau) \tag{13}$$

where $N(t)$ is a rxr symmetric positive definite matrix.

where e) the state vector $x(t)$ at initial time t_o is a Gaussian process with known mean

$$E\{x(t_o)\} \triangleq \bar{x}_o \tag{14}$$

and covariance

$$E\{(x(t_o) - \bar{x}_o)\ \ (x(t_o) - \bar{x}_o)^T\} \triangleq P_o \tag{15}$$

where f) the measurement noise and the system noise are statistically independent, that is

$$E\{w(t)v(t)^T\} = 0 \tag{16}$$

where g) the initial state vector and the system noise are statistically independent

$$E\{x(t_o)w^T(t)\} = 0 \text{ for all } t \geq t_o. \tag{17}$$

The Kalman-Bucy filter is a differencing filter. It is represented by the following stochastic differential equation

$$\dot{\hat{x}}(t) = f(t)\hat{x}(t) + B(t)u(t) + L(t)y(t) \tag{18}$$

$$\hat{x}(t_o) = \hat{x}_o = \bar{x}_o \tag{19}$$

It is the objective of the filter to produce the "best" unbiased estimate of the state of the system. To achieve the "best" estimate the covariance of the error, where the error is defined to be

$$e(t) = x(t) - \hat{x}(t) \tag{20}$$

must be minimized for all time and thus a minimum at the terminal time, T. This can be accomplished by optimizing the filter using a cost functional of the following form

$$J = E\left\{e^T(t)e(t)\right\} \tag{21}$$

where $T \geq t_o$. The above cost function satisfies the criterion necessary in using the stochastic maximum principle. Applying the trace identity and the definition of the covariance to equation (21) the cost function becomes

$$\begin{aligned} J &= E\left\{ tr(e(t)e^T(t)) \right\} \\ &= tr\, E\left\{ e(t)e^T(t) \right\} \\ &= tr\, P(t) \end{aligned} \tag{22}$$

where $P(t)$ is the covariance of the error. Before deriving the vector differential equation for the error covariance it is necessary to derive the equation for the error. This is accomplished by differentiating the error equation (20)

$$\dot{e}(t) = \dot{x}(t) - \dot{\hat{x}}(t). \tag{23}$$

Using equation (23) and the equations for the system (7) and the filter (18) the error equation becomes

$$\begin{aligned} \dot{e}(t) = &\left\{ A(t) - F(t) - L(t)H(t) \right\} x(t) + F(t)e(t) + G(t)w(t) \\ &- L(t)v(t). \end{aligned} \tag{24}$$

Applying the condition that the output of the optimal Kalman-Bucy filter be an unbiased estimate of the state of the system results in

$$E\left\{ x(t) \right\} = E\left\{ \hat{x}(t) \right\} \text{ for all } t \geq t_o. \tag{25}$$

This constraint imposes the following condition on the error

$$E\left\{ e(t) \right\} = E\left\{ x(t) - \hat{x}(t) \right\} = 0 \text{ for all } t \geq t_o. \tag{26}$$

Since

$$\frac{d}{dt} E\left\{ e(t) \right\} = E\left\{ \dot{e}(t) \right\} \tag{27}$$

then

$$E\left\{ e(t) \right\} = 0 \text{ for all } t \geq t_o. \tag{28}$$

Taking the expected value of error differential equation (24) and substituting equations (28), (25), (10), (12), and (26) into it yields

$$0 = \left\{ A(t) - F(t) - L(t)H(t) \right\} E\left\{ x(t) \right\} \text{ for all } t \geq t_o. \tag{29}$$

The non-trivial solution for equation (29) results in

$$F(t) = A(t) - L(t)H(t) \tag{30}$$

which specifies one of the two unknown matrices of the filter equations. Substituting this result into the differential equation for the error (24) yields

$$\dot{e}(t) = A(t) - L(t)H(t)\, e(t) + G(t)w(t) - L(t)v(t). \tag{31}$$

The covariance equation for the error is given by the following differential equation and is derived in (7).

$$\begin{aligned} \dot{P}(t) = &\left[A(t) - L(t)H(t) \right] P(t) + P(t) \left[A(t) - L(t)H(t) \right]^T \\ &+ G(t)Q(t)G^T(t) + L(t)N(t)L^T(t) \end{aligned} \tag{32}$$

The initial condition for this error covariance equation is

$$\begin{aligned} P(t_o) &= E\left\{e(t_o)e^T(t_o)\right\} \\ &= E\left\{\left[x(t_o) - \hat{\bar{x}}(t_o)\right]\left[x(t_o) - \hat{\bar{x}}(t_o)\right]^T\right\} \\ &= E\left\{\left[x(t_o) - \bar{x}(t_o)\right]\left[x(t_o) - \bar{x}(t_o)\right]^T\right\} \\ &= P_o. \end{aligned} \tag{33}$$

The remaining unknown quantity of the filter is the optimal admissible gain matrix $L^*(t)$. Once evaluated, the error covariance equation will have completely specified elements and the cost functional will be minimized. Let $L(t)$ be any admissible gain matrix for $t\varepsilon\left[t_o,T\right]$. The object of the following optimization procedure will be to select from all the possible admissible gain matrices that one which minimizes the cost functional at the terminal time, T. In order to derive this optimal admissible gain matrix $L^*(t)$, it becomes necessary to introduce the costate matrix $\lambda(t)$. The costate matrix $\lambda(t)$ is defined such that each element of the error covariance matrix $P(t)$ has assigned to it an element of the costate matrix. In doing so, the Hamiltonian equation becomes

$$H(P(t),L(t),\lambda(t)) = tr\left[\lambda^T(t)\dot{P}(t)\right]. \tag{34}$$

If the elements of the error covariance matrix $P(t)$ are viewed as state variables and the elements of the gain matrix as the control variables the cost functional (22) will be optimized by the selection of the optimal gain matrix $L^*(t)$. Substituting the error covariance equation (32) into the Hamiltonian equation (34) yields

$$\begin{aligned} H(P(t),L(t),\lambda(t)) = tr\Big[&\lambda^T(t)A(t)P(t) - \lambda^T(t)L(t)H(t)P(t) \\ &+ \lambda^T(t)P(t)A^T(t) - \lambda^T(t)P(t)H^T(t)L^T(t) + \lambda^T(t)G(t)Q(t)G^T(t) \\ &+ \lambda^T(t)L(t)N(t)L^T(t)\Big]. \end{aligned} \tag{35}$$

The stochastic maximum principle requires that for this application the following equation must hold

$$E\left\{H(P(t),L^*(t),\lambda(t)) \mid y(t)\right\} \leq E\left\{H(P(t),L(t),\lambda(t)) \mid y(t)\right\}. \tag{36}$$

where $L^*(t)$ symbolizes the optimal admissible gain matrix. In addition the stochastic maximum principle has as a necessary condition that

$$\left.\frac{\partial E\left\{H(P(t),L(t),\lambda(t)) \mid y(t)\right\}}{\partial L(t)}\right|_* = 0 \tag{37}$$

where (*) denotes that the expression is evaluated along the optimal trajectory. Solving equation (37) for the optimal admissible gain matrix yields

$$\begin{aligned} &- P(t)H^T(t)E\left\{\lambda(t) \mid y(t)\right\} - E\left\{\lambda^T(t) \mid y(t)\right\}P(t)H^T(t) \\ &+ E\left\{\lambda(t) \mid y(t)\right\}L^*(t)N^T(t) + E\left\{\lambda^T(t) \mid y(t)\right\}L^*(t)N(t) = 0 \end{aligned} \tag{38}$$

Since $\lambda(t)$ is symmetric positive definite and $N(t)$ is assumed symmetric positive definite then $\lambda(t) = \lambda^T(t)$ and $N(t) = N^T(t)$. Equation (38) can be rewritten in the form

$$E\left\{\lambda(t) \mid y(t)\right\}\left[- P(t)H^T(t) - P(t)H^T(t) + L^*(t)N(t) + L^*(t)N(t)\right] = 0. \tag{39}$$

The non-trivial solution of equation (39) is

$$- 2P(t)H^T(t) + 2L^*(t)N(t) = 0. \tag{40}$$

Solving equation (40) for the optimal gain matrix $L^*(t)$ yields

$$L^*(t) = P(t)H^T(t)N^{-1}(t). \tag{41}$$

Substituting the optimal admissible gain matrix $L^*(t)$ into the error covariance equation (32) and the optimal filter equation (18) yields the well-known Kalman-Bucy filter equations

$$\dot{\hat{x}}(t) = A(t)\hat{x}(t) + B(t)u(t) + P(t)H^T(t)R^{-1}(t)\left[y(t) - H(t)\hat{x}(t)\right] \tag{42}$$

$$\hat{x}(t_o) = \hat{x}_o \tag{43}$$

$$\dot{P}(t) = A(t)P(t) + P(t)A^T(t) - P(t)H^T(t)R^{-1}(t)H(t)P(t) + G(t)Q(t)G^T(t), \; P(t_o) = P_o \tag{44}$$

FORMULATION OF THE OPTIMAL CONTROL

To complete the formulation of the optimal controller it is necessary to derive the equations describing the optimal deterministic controller. Once derived, the optimal deterministic controller is cascaded with the optimal Kalman-Bucy filter to form the optimal controller which optimizes the system using a quadratic cost functional.

The optimal control problem can be stated as follows; given

(1) the system represented by the equations

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + G(t)w(t) \tag{45}$$

$$x(t_o) = x_o \tag{46}$$

$$y(t) = H(t)x(t) + v(t) \tag{47}$$

(2) the optimal Kalman-Bucy filter represented by the equations

$$\dot{\hat{x}}(t) = A(t)\hat{x}(t) + B(t)u(t) + P(t)H^T(t)R^{-1}(t)\; y(t) - H(t)\hat{x}(t) \tag{48}$$

$$\hat{x}(t_o) = \hat{x}_o = \bar{x}_o \tag{49}$$

$$\dot{P}(t) = A(t)P(t) + P(t)A^T(t) - P(t)H^T(t)N^{-1}(t)H(t)P(t) + G(t)Q(t)G^T(t), \; P(t_o) = P_o \tag{50}$$

and (3) the quadratic cost functional

$$J = E\left\{\int_o^T (x^T(t)S(t)x(t) + u^T(t)R(t)u(t))dt\right\} \tag{51}$$

where T is a fixed time, S(t) is an nxn positive semidefinite matrix and R(t) is an mxm positive definite symmetric matrix; find the optimal control equations using the stochastic maximum principle.

The state variables, $x(t)$, of the system are corrupted by additive white Gaussian noise as are the observation variables, $y(t)$. It is the object of the optimal filter to obtain the "best" unbiased estimate of the state variables given the imperfect observations. Since the optimal filter can not perfectly estimate the state variables of the system and error term exists and is represented by

$$e(t) = x(t) - \hat{x}(t) \tag{52}$$

where $\hat{x}(t) = E\{x(t) \mid y(t)\}$ is the optimal unbiased estimate of $x(t)$. Since the error differential equation (31) is independent of the admissible control vector $u(t)$ and since

$$\operatorname{cov}\left[e(t);\ \hat{x}(t)\right] = 0 \tag{53}$$

then the cost functional can be written in terms of the optimal estimation and the admissible control vector $u(t)$,

$$J = E\left\{\int_0^T \hat{x}^T(t)S(t)\hat{x}(t) + u^T(t)R(t)u(t)\ dt\right\}. \tag{54}$$

The Hamiltonian equation associated with the cost functional (54) is

$$H(\hat{x}(t),u(t),\lambda(t)) = \lambda^T(t)f(\hat{x}(t),u(t)) + x^T(t)S(t)x(t) + u^T(t)R(t)u(t) \tag{55}$$

where

$$\dot{\hat{x}}(t) = f(\hat{x}(t),u(t)) = A(t)\hat{x}(t) + B(t)u(t) + P(t)H^T(t)N^{-1}(t) \cdot y(t) - H(t)\hat{x}(t)\ . \tag{56}$$

The stochastic maximum principle when applied to this problem is given by

$$E\left\{H(\hat{x}(t),u^*(t),\lambda(t)) \mid \hat{x}(t)\right\} \le E\left\{H(\hat{x}(t),u(t),\lambda(t)) \mid \hat{x}(t)\right\} \tag{57}$$

where $u(t)$ is any admissible control and $u^*(t)$ is the optimal admissible control. In addition, the optimal admissible control must satisfy the following equation

$$\frac{\partial\, E\left\{H(\hat{x}(t),u(t),\lambda(t)) \mid \hat{x}(t)\right\}}{\partial u(t)}\ \Bigg|_{*} = 0 \tag{58}$$

where $(*)$ denotes that the expression is to be evaluated along the optimal trajectory. Solving equation (58) yields

$$B^T(t)E\left\{\lambda^T(t) \mid \hat{x}(t)\right\} + 2R(t)u^*(t)\Big\} = 0 \tag{59}$$

The optimal admissible control is derived from equation (59) and is

$$u^*(t) = -\,1/2\ R^{-1}(t)B^T(t)E\left\{\lambda^T(t) \mid \hat{x}(t)\right\}. \tag{60}$$

A particular solution for the costate matrix $\lambda(t)$ is

$$\lambda(t) = 2K(t)\hat{x}(t). \tag{61}$$

Taking the conditional expectation of the costate equation (61) yields

$$E\left\{\lambda(t) \mid \hat{x}(t)\right\} = 2K(t)\hat{x}(t). \tag{62}$$

Substituting equation (62) into the equation for the optimal admissible control $u^*(t)$ yields

$$u^*(t) = -R^{-1}(t)B^T(t)K(t)\hat{x}(t). \tag{63}$$

All the terms in equation (63) are defined except the optimal control gain matrix K(t). To evaluate the optimal control gain matrix K(t), the differential equation of the costate equation must be evaluated and set equal to the time derivative of equation (61). The relationship for the optimal admissible control was evaluated using the conditional expectation of the Hamiltonian equation. Therefore, when evaluating the costate differential equations it is necessary to take the condition expectation of the equations. Taking the conditional expectation of equation (5) yields

$$E\left\{\lambda(t) \mid \hat{x}(t)\right\} = \frac{\partial E\left\{H(\hat{x}(t), u(t), \lambda(t)) \mid \hat{x}(t)\right\}}{\partial \hat{x}(t)} = A^T(t)\lambda(t) + 2S(t)\hat{x}(t). \tag{64}$$

The conditional expectation of the time derivative of equation (61) is

$$E\left\{\lambda(t) \mid \hat{x}(t)\right\} = 2\dot{K}(t)\hat{x}(t) + 2K(t)E\left\{\dot{\hat{x}}(t) \mid x(t)\right\}. \tag{65}$$

Setting equations (64) equal to equation (65) and substituting in the equations for the optimal admissible control and the filter equation yields

$$\left\{\dot{K}(t) + A^T(t)K(t) + K(t)A(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) + S(t)\right\}\hat{x}(t) = 0 \tag{66}$$

The non-trivial solution of equation (66) yields the optimal control gain differential equation

$$-\dot{K}(t) = K(t)A(t) + A^T(t)K(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) + S(t). \tag{67}$$

Using the solution for the costate vector (61) and equation (4) the final condition for equation (67) becomes

$$K(T) = 0 \tag{68}$$

SIMULATION

The theory for optimizing an n^{th} order continuous stochastic process and its application to a generalized n^{th} order linear continuous stochastic process was presented in the previous sections. In this section the generalized n^{th} order continuous linear time-invariant optimal stochastic process with imperfect measurements is simulated on the hybrid computer. The hybrid computer used for this simulation is an Electronic Associates, Inc. EAI-8900 hybrid computer.

The hybrid analog computer is used to simulate (1) the continuous linear time-invariant system which is corrupted by system and measurement noise, (2) the Kalman-Bucy filter and (3) the optimal deterministic type controller. The equations for the above system are

(1) the continuous linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t) + Gw(t) \tag{69}$$

$$x(t_o) = x_o \tag{70}$$

$$y(t) = Hx(t) + v(t) \tag{71}$$

(2) the optimal Kalman-Bucy filter

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + P(t)H^T N^{-1}(y(t) - H\hat{x}(t)) \tag{72}$$

$$\hat{x}(t_o) = \bar{x}_o \tag{73}$$

(3) the optimal deterministic controller

$$u^*(t) = - B^T R^{-1} K(t)\hat{x}(t) \tag{74}$$

These equations are programmed on the computer in their most generalized form so as to provide for every possible combination of the system parameters.

The digital computer assumes absolute control over the hybrid simulation, generates the noise sources for the simulation and generates the solutions to the nonlinear matrix Riccati equations. These equations are

(1) the optimal controller Riccati equation

$$-\dot{K}(t) = A^T K(t) + K(t)A - K(t)BR^{-1}B^T K(t) + S(t) \tag{75}$$

$$K(T) = 0 \tag{76}$$

(2) the Kalman-Bucy filter Riccati equation

$$\dot{P}(t) = P(t)A^T + AP(t) - P(t)H^T N^{-1} HP(t) + GQG^T \tag{77}$$

$$P(t_o) = 0 \tag{78}$$

The problem simulated for this paper is the "Broom Balancing Problem" described in O. Elgerd book, Control System Theory, McGraw Hill, Inc., 1967. The equations of motion for this problem, in state variable form are

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{g(m+M)mL}{I(m+M)+mML^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{gm^2L^2}{I(m+M)+mML^2} & 0 & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{-mL}{I(m+M)+mML^2} \\ 0 \\ \frac{I+mL^2}{I(m+M)+mML^2} \end{bmatrix} u(t)$$

$$+ \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix} \tag{79}$$

$$y(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} \tag{80}$$

where g = gravitational constant in meters per seconds square

m = mass of the broom in kilograms

M = mass of the cart in kilograms

L = length of the broom in meters

I = moment of inertia of the broom

x_1 = angular displacement of the broom in degrees

x_2 = angular velocity of the broom in degrees per second

x_3 = cart displacement in meters

x_4 = velocity of the cart in meters per second

$w_1(t), w_2(t)$ = random wind vectors

$v_1(t), v_2(t)$ = measurement noise vectors.

For this simulation the system parameters are

$$m = M = 1.00 \text{ kg} \qquad L = 1.00 \text{ meters} \qquad I = 0.3333 \text{ m}^2\text{-kg}$$

$$g = 10.0 \text{ m/sec}^2 \qquad x_1(t_o) = 0.5 \text{ degrees}$$

$$\text{cov}\left[w(t);w(\tau)\right] = \text{cov}\left[v(t);v(\tau)\right] = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \delta(t-\tau)$$

$$R = \begin{bmatrix} 1.0 \end{bmatrix} \qquad S = \begin{bmatrix} 10.0 & 0.0 \\ 0.0 & 10.0 \end{bmatrix}$$

The series of recordings shown in Figure 1 demonstrates that the optimal control law achieves its objectives. The inherently unstable system is stabilized and the angular displacement of the broom and the translational displacement of the cart are forced to oscillate about a zero value.

CONCLUSIONS

The results of Kushner's stochastic maximum principle were presented and applied to an n^{th} order continuous linear time-varying system subject to plant and measurement noise. This system meets the three requirements of the separation principle thus enabling the controller to be constructed by cascading an optimal deterministic controller and an optimal estimator. The derivation of the optimal estimator and controller is verified by existing results derived by more conventional methods.

To demonstrate the theory, a fourth order system was simulated on the hybrid computer and sample recordings are given in the text.

REFERENCES

[1] Kushner, H. J., "On Stochastic Extremum Problems: Calculus". Journal of Math Analysis and Appl., Vol. 10, No. 2, April 1965, pp. 354-367.

[2] Kushner, H. J., "On the Existence of Optimal Stochastic Controls". Journal SIAM Control, Series A, Vol. 3, No. 3, 1966, pp 463-474.

[3] Kushner, H. J. and Schweppe, F. C., "A Maximum Principle for Stochastic Control Systems". Journal of Math Analysis and Appl.,Vol. 8, 1964, pp. 287-302.

[4] Kushner, H. J., "Sufficient Conditions for the Optimality of a Stochastic control". Journal SIAM Control, Series A, Vol. 3, No. 3, 1966, pp. 499-508.

[5] Sworder, D. D., "On the Control of Stochastic Systems". International Journal of Control, Vol. 6, No. 2, pp. 179-188, 1967.

[6] Sworder, D. D., "Feedback Control of a Class of Linear Stochastic Systems". Ninth Joint Automatic Control Conference Preprints, University of Michigan, Ann Arbor, Michigan, June, 1968.

[7] Athans, M. and Tse, E., "A Direct Derivation of the Optimal Linear Filter Using the Maximum Principle". IEEE Transactions on Automatic Control, Vol. AC-12, No. 6, pp. 690-698, 1967.

[8] Kushner, H. J., "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes, with Applications". J. SIAM Control, Series A, Vol. 2, No. 1, 1964.

[9] Kushner, H. J., "On the Dynamical Equations of Conditional Probability Density Functions, with Applications to Optimal Stochastic Control Theory". Journal of Math Analysis and Appl. Vol. 8, No. , pp. 332-344, 1964.

[10] Kushner, H. J., "On the Stochastic Maximum Principle with "Average" Constraints". Journal of Mathematical Analysis and Applications, Vol. 12, No. 1, August 1965, pp. 13-26.

[11] Kushner, H. J., "On the Stochastic Maximum Principle with Fixed Time of Control". Journal of Math Analysis and Appl., 11, pp. 78-92, (1965).

[12] Kushner, H. J., "Some Problems and Some Recent Results in Stochastic Control". IEEE Int'l. Convention Record, 1965, Part 6, pp.108-116.

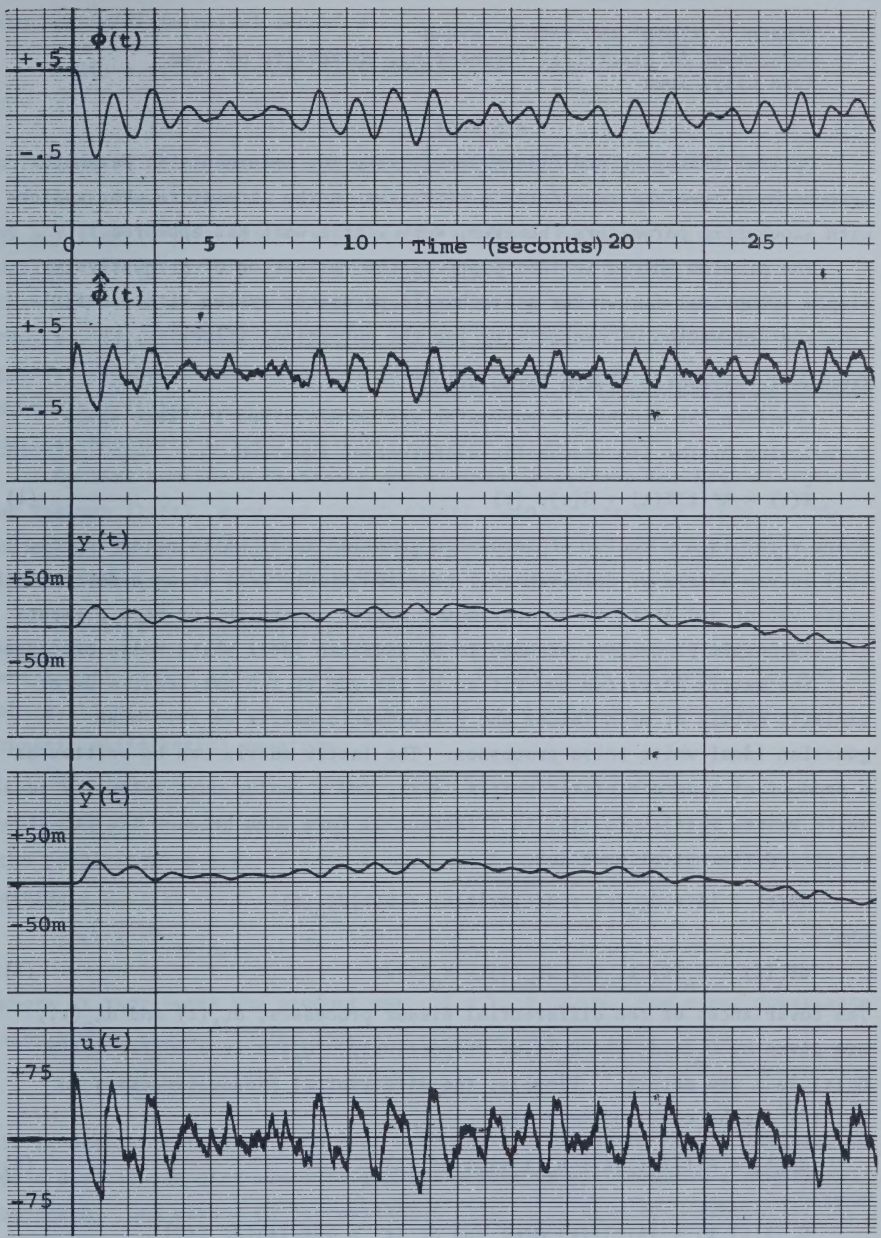


Figure 1

A SUBOPTIMAL FILTER AND FILTER SENSITIVITY FOR STOCHASTIC PARAMETER SYSTEMS

RONALD L. KLEIN* AND EARL D. EYMAN
Department of Electrical Engineering, University of Iowa
Iowa City, Iowa

ABSTRACT

Optimal filtering and filter sensitivity for deterministic parameter systems have received considerable attention in the literature. For systems with stochastic parameters, however, very little work has appeared.

The problem solved in this paper is the derivation of a relatively simple suboptimal filter for a stochastic parameter system with a definition and derivation of sensitivity for the suboptimal filter.

Stochastic Parameter System Formulation

Consider the stochastic parameter system described in state space formulation by the Ito stochastic differential equation

$$\dot{X}(t) = A^*(t)X(t) + G(t)v_o(t) \tag{1}$$

where $v_o(t)$ is a p vector of zero-mean gaussian ideal white noise inputs, G(t) is nxp matrix of deterministic time functions, X(t) is an n-dimensional state vector and A*(t) is the parameter matrix, where one or more elements are stochastic processes, assumed to be gaussian ideal white noise processes.

The matrix A*(t) can be written as the sum of two matrices; a deterministic matrix of mean values and a stochastic matrix of zero-mean gaussian ideal white noise processes. The latter matrix can be written as a product of a deterministic matrix, $\bar{A}(t)$, and a scalar unit intensity ideal white noise, $v_A(t)$. Eq. (1) in differential form then becomes the Ito stochastic differential equation

$$dX(t) = [\bar{A}(t)d\xi_A(t) + A(t)dt]X(t) + G(t)d\xi_o(t) \tag{2}$$

where

$$d\xi_A(t) = v_A(t)dt, \text{ and } d\xi_o(t) = v_o(t)dt \tag{3}$$

The covariances of the differential Wiener processes, $d\xi_A(t)$ and $d\xi_o(t)$, are given by $Q_A(t)dt$ and $Q_o(t)dt$ respectively. $Q_A(t)$ is a scalar while $Q_o(t)$ is a pxp matrix. Finally, the differential Wiener processes in Eq. (2) can be combined to yield

$$dX(t) = A(t)X(t)dt + G_t(X(t),t)d\xi_t(t) \tag{4}$$

where

$$G_t(X(t),t) = [\bar{A}(t)X(t) : G(t)] \text{ and } d\xi_t(t) = [d\xi_A(t) : d\xi_o(t)]^T.$$

* Now with the Department of Electrical Engineering, Kansas University, Lawrence, Kansas.

Equation (4) is recognized as a special case of the general nonlinear stochastic differential equation

$$dX(t) = f(X(t),t)dt + G(X(t),t)d\xi(t),$$
$$E[d\xi(t)d\xi^T(t)] = Q(t)dt \tag{5}$$

solutions of which are well known to be diffusive Markov processes.

With the stochastic parameter system given by Eq. (4), the derivation of a filter for the system proceeds as follows. First, a derivation of the optimal filter for the general nonlinear system of Eq. (5) is outlined, then the resulting filter is specialized to the stochastic parameter system, Eq. (4), to find the optimal filter for this system. Finally the complexity of the optimal filter is simplified by defining a satisfactory suboptimal filter.

Optimal Nonlinear Filter

Optimal nonlinear filtering for stochastic systems has been discussed by Kushner (1964), Bucy and Joseph (1968), and Stratonovich (1960). The objective is minimization of the conditional expected value of the weighted error; that is, the system cost function has the form

$$J = <[X(t) - \hat{X}(t)]W(t)\ [X(t) - \hat{X}(t)]^T | Y> \tag{6}$$

where $X(t)$ is the system state, and $\hat{X}(t)$ is the state estimate. The set of observations, $Y = \{y(\tau): \tau \leq t\}$ is associated with the measurement stochastic differential equation

$$dy(t) = H(X(t),t)dt + d\eta(t)\ ,\ E[d\eta(t)d\eta^T(t)] = R(t)dt \tag{7}$$

The matrix $W(t)$ is a weighting matrix and $< >$ denotes expected value. If the conditional p.d.f. of $X(t)$ given Y is denoted by $p(x, t|Y, t)$ then the estimate $\hat{X}(t)$ is

$$\hat{X}(t) = \int_{-\infty}^{\infty} xp(x,t|Y,t)dx \tag{8}$$

where dx represents the product πdx_i, $i = 1, 2, \ldots n$. The error covariance matrix with respect to the same distribution can be shown to be

$$P(t) = \int_{-\infty}^{\infty} xx^T p(x,t|Y,t)dx - \hat{X}(t)\hat{X}^T(t). \tag{9}$$

It is desirable to have a form of Eqs. (8) and (9) which can be implemented on a digital computer since in general solutions will require a numerical approach. One method of accomplishing this is to develop differential equations which can be solved numerically. Taking differentials of Eqs. (8) and (9) gives

$$d\hat{X}(t) = \int_{-\infty}^{\infty} x\,dp(x,t|Y,t)dx \tag{10}$$

$$dP(t) = \int_{-\infty}^{\infty} xx^T dp(x,t)dx - d[\hat{X}(t)\hat{X}^T(t)]. \tag{11}$$

The differentials in (10) and (11) are interpreted as the differentials resulting from a differential increase in time, dt.

To simplify the above expressions for the state estimate and error covariance, an expression for $dp(x,t|Y,t)$, the differential conditional p.d.f., is needed. The derivation of this expression is too lengthy to be included here; however an overview of the steps involved follows.

First an arbitrary twice differentiable function, F, of the state, X(t), is assumed and the differential, dF, of F is calculated using Ito's lemma. Then dF is integrated over an incremental time interval, δt, and the expected value of the result with respect to $p(x_{t+\delta t}, t+\delta t|Y,t;\delta y,\delta t)$ is taken. Finally allowing δt to approach zero in the limit yields

$$dp(x,t|Y,t) = \left(\frac{\partial}{\partial x}\right)^T [f(x,t)dt + G(x,t)C(t)R^{-1}(t)(dy-\hat{H}(x,t)dt)]$$
$$\cdot\, p(x,t|Y,t) + \frac{1}{2}\,\mathrm{tr}\left(\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial x}\right)^T [G(x,t)Q(t)G^T(x,t)]dt$$
$$\cdot\, p(x,t|Y,t) + [H(x,t)-\hat{H}(x,t)]^T R^{-1}(t)[dy-\hat{H}(x,t)dt]p(x,t|Y,t). \tag{12}$$

Using this result in Eqs. (10) and (11) and using Ito's lemma to compute the second term on the right hand side of (11) results in the following nonlinear filter differential equations,

$$d\hat{X}(t) = \hat{f}(X,t)dt + [\langle XH^T(X,t)|Y\rangle - \hat{X}(t)\hat{H}^T(X,t)$$
$$+ \hat{G}(X,t)C(t)]R^{-1}(t)(dy-\hat{H}(X,t)dt)\ ;\ \hat{X}(t_o) = \hat{X}_o. \tag{13}$$

$$dP(t) = \langle(X-\hat{X})f^T\rangle dt + \langle f(X-\hat{X})^T\rangle dt + \langle(X-\hat{X})\,[GCR^{-1}(dy-\hat{H}dt)]^T\rangle$$
$$+ \langle GCR^{-1}(dy-\hat{H}dt)(X-\hat{X})^T\rangle + \langle GQR^T\rangle dt$$
$$- [\langle XH^T\rangle - \hat{X}\hat{H}^T + \hat{G}C]R^{-1}[\langle HX^T\rangle - \hat{H}\hat{X}^T + C^T\hat{G}^T]dt$$
$$+ \langle(X-\hat{X})(X-\hat{X})^T(H-\hat{H})^T R^{-1}(dy-\hat{H}dt)\rangle\ ,\ P(t_o) = P_o \tag{14}$$

These differential equations are to be interpreted in the Ito sense and apply to any nonlinear system of the form (5) given the set of measurements Y from (7).

Optimal Filter Equations for the Stochastic Parameter System

To determine the optimal filter equations for the stochastic parameter system, Eqs. (13) and (14) are applied to Eq. (4) with measurements assumed in the form

$$dy(t) = H(t)X(t)dt + d\eta(t),\ y(t_o) = y_o \tag{15}$$

or

$$Z(t) = \dot{y}(t) = H(t)X(t) + \dot{\eta}(t),\ Z(t_o) = 0 \tag{16}$$

Correlation between the $\eta(t)$ measurement noise and the total system noise is

$$\langle d\xi_t(t)d\eta^T(t)\rangle = C(t)dt = [C_1(t) \vdots C_2(t)]^T dt \tag{17}$$

where $C_1(t)$ is taken to be zero based upon the assumption that parameters and measurement noise are uncorrelated. Measurement noise is assumed to be correlated as in Eq. (7). Finally the parameters and the system noise are assumed uncorrelated so that

$$\langle d\xi_t(t)d\xi_t^T(t)\rangle = Q_t(t)dt = \begin{bmatrix} \vdots & \\ 1 \vdots & 0 \\ \cdots\cdots & \\ 0 \vdots & Q_o(t) \\ \vdots & \end{bmatrix} dt. \tag{18}$$

When the vector function, $f(X(t),t)$, and matrices $G(X(t),t)$ and $H(X(t),t)$ of Eqs. (5) and (7) are identified for the stochastic parameter system, and the expected values are taken and substituted into Eq. (13), the optimal estimate differential equation for the stochastic parameter system becomes

$$d\hat{X} = A\hat{X}dt + [PH^T + GC_2]R^{-1}(dy - H\hat{X}dt). \tag{19}$$

Applying a similar procedure to the error covariance differential equation yields

$$\begin{aligned} dP = {} & [PA^T + AP + \tilde{A}P\tilde{A}^T + \tilde{A}\hat{X}\hat{X}^T\tilde{A}^T + GQ_oG^T - PH^TR^{-1}HP \\ & - PH^TR^{-1}C_2^TG^T - GC_2R^{-1}HP - GC_2R^{-1}C_2^TG^T]dt \\ & + \langle (X-\hat{X})(X-\hat{X})^T(X-\hat{X})^TH^TR^{-1}(dy-H\hat{X}dt)\,|Y\rangle. \end{aligned} \tag{20}$$

It is seen by inspection that this differential equation depends upon the third order moment of the error $(X-\hat{X})$; a characteristic of systems in which the G matrix of Eq. (5) (for the stochastic parameter system, the G_t matrix of Eq. (4)) depends upon the state, X(t). The parameters are the source of this dependence in the stochastic parameter system. As a result of the third order error dependence the error does not have a precise gaussian distribution. It is important to note, however, that the third order term does not exhibit direct dependence on the parameter matrix $\tilde{A}$, the matrix controlling the stochastic part of the parameters. Using this hint, the term involving the third order error moment is neglected and the error distribution is approximated by its gaussian part, or the first two moments.

Neglecting the term does, however, result in a suboptimal filter.

Equation (20) simplifies to

$$\dot{P} = P[A - GC_2R^{-1}H]^T + [A - GC_2R^{-1}H]P - PH^TR^{-1}HP + \tilde{A}P\tilde{A}^T + \tilde{A}\hat{X}\hat{X}^T\tilde{A}^T + G(Q_o - C_2R^{-1}C_2^T)G^T \ , \ P(t_o) = P_o. \tag{21}$$

A comparison of the filtering equations (19) and (21) with the corresponding well-known Kalman-Bucy filter equations shows that 1) the optimal estimate differential equation is identical and 2) the differential equation for the covariance matrix, P(t), is no longer of the Riccati type, independent of the state estimate $\hat{X}$, but now depends on the state estimate and also on the parameter covariance matrix $\tilde{A}$. P(t) can be calculated a priori in Kalman-Bucy filtering. For the case at hand, due to the coupling with $\hat{X}(t)$ and $\tilde{A}(t)$, the P(t) differential equation must be solved simultaneously with the $\hat{X}(t)$ differential equation.

Sensitivity of the Suboptimal Filter

The suboptimal filter sensitivity for stochastic parameter systems is defined in terms of the error covariance of the filter estimate. A sensitivity for Kalman-Bucy filters applied to deterministic parameter systems was defined in a similar way by Griffin and Sage (1968). Using the formulation of a stochastic parameter system, Eq. (4), the sensitivity is defined as the partial derivative of P(t), the filter error covariance, with respect to $\tilde{a}_{ij}(t)$, that is

$$S_{ij}(t) = \frac{\partial P(t)}{\partial a_{ij}} . \tag{22}$$

Sensitivity differential equations can be calculated using the definition, (22), and the covariance matrix differential equation, (21),

$$\dot{S}_{ij}(t) = \frac{d}{dt}\left[\frac{\partial P(t)}{\partial \tilde{a}_{ij}(t)}\right] = \frac{\partial}{\partial \tilde{a}_{ij}}[\dot{P}(t)] \tag{23}$$

assuming the order of differentiation can be interchanged. The resulting differential equation is

$$\dot{S}_{ij} = S_{ij}[A - (GC_2 + PH^T)R^{-1}H]^T + [A - (GC_2 + PH^T)R^{-1}H]S_{ij} + [1_{ij}][P + \hat{X}\hat{X}^T]\tilde{A}^T + \tilde{A}[P + \hat{X}\hat{X}^T][1_{ji}] + \tilde{A}S_{ij}\tilde{A}^T. \tag{24}$$

where $[1_{ij}]$ and $[1_{ji}]$ are zero matrices excepting the 1 in the indicated positions. The initial condition for (24) is found by differentiating the initial error covariance with respect to $\tilde{a}_{ij}(t_o)$

$$S_{ij}(t_o) = \frac{\partial P(t_o)}{\partial \tilde{a}_{ij}(t_o)} . \tag{25}$$

A Fortran IV simulation program was written which solves the differential equations for the stochastic parameter system, Eq. (4), the filter estimate, Eq. (19), the error covariance, Eq. (21), the sensitivities of the form of Eq. (24), and computes the measurements according to Eq. (16). Numerical integration was accomplished by using a fourth order Runge-Kutta method proposed by Gill (1951). The program was used to simulate two stochastic parameter systems; a first order and a second order system. The first order system is described by

$$\dot{X}(t) = [0.2\, u_A(t) - 1]X(t) + u_o(t) \tag{26}$$

where $u_A(t)$ and $u_o(t)$ are physical white gaussian processes corresponding to the idealized white noise processes $v_A(t)$ and $v_o(t)$ of equation (3). Measurements were of the form

$$Z(t) = X(t) + v_m(t) \tag{27}$$

Power levels for the processes $u_A(t)$ and $u_o(t)$ were 2mw/HZ and 18mw/HZ respectively. The physical measurement noise, $v_m(t)$, corresponding to $\dot{n}(t)$ in Eq. (16), was identical to $v_o(t)$ so that the system and measurement disturbances are unity correlated and have identical power levels.

The solutions for the system state and filter estimate of the state are plotted in Fig. 1 for a time interval of approximately 4 seconds. Figure 2 gives the measurements of the system taken according to Eq. (16). It is interesting to note the extremely random appearance of the measurement data upon which the filter operates to produce the high quality state estimate given in Figure 1. Figures 3 and 4 give respectively the error covariance and sensitivity solutions calculated by using Eqs. (21) and (24) applied to the first order system.

A second order system described by

$$\dot{X}(t) = \left[\begin{bmatrix} .1 & .1 \\ .1 & .1 \end{bmatrix} u_A(t) + \begin{bmatrix} -1 & 2 \\ 0 & -2 \end{bmatrix}\right] X(t) + \begin{bmatrix} .5 & .25 \\ .25 & .5 \end{bmatrix} u_o(t) \tag{28}$$

with measurements

$$Z(t) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} X(t) + v_m(t)\ , \tag{29}$$

was also simulated with power levels of all four parameters equal to 0.5mw/HZ. The covariance matrix, Q(t), of the system disturbance processes, $u_o(t)$, was chosen to have diagonal elements equal to 0.5 and off diagonal elements equal to 0.25. As in the first order system, the system and measurement disturbances were identical, therefore $R(t) = C_2(t) = Q(t)$. Power levels of each of the element processes of $u_o(t)$ was 25mw/HZ.

Figures 5 through 12 record the computational results for this system. The convention used in the sensitivity subscripts is that the first pair indicate the error covariance element appearing in the numerator of the sensitivity while the second pair of subscripts indicate the element of the $\tilde{A}(t)$ matrix which appears in the denominator of the sensitivity. Sensitivity matrix elements for $P_{11}(t)$ and $P_{12}(t)$ are given in Figures 11 and 12 respectively.

Results

The suboptimal filter applied to stochastic parameter systems was found to be computationally stable for the systems simulated as demonstrated by the Figures. Further, the filter estimate rapidly converges toward and tracks the actual state very accurately as demonstrated by Figures 1, 5 and 6 which show the actual state and the state estimate.

The sensitivity, as defined, offers easy comparison between the stochastic parameters with respect to contributions made to the estimate error. For example, mere inspection of Figures 9 and 10 shows that the $\tilde{a}_{11}(t)$ element of $\tilde{A}(t)$ affects $P_{11}(t)$ more drastically than any other element.

Application of the sensitivity analysis illustrated in this paper to the determination of optimally sensitive stochastic parameter systems offers an interesting continuation of the above research.

In summary, a suboptimal filter for a stochastic parameter system with additive disturbances has been derived and demonstrated. Further, a sensitivity formulation has been derived which relates the filter estimate error and the stochastic parameters. System simulations demonstrate two practical applications of the results.

REFERENCES

Bucy, R. S., and Joseph, P. D., Filtering for Stochastic Processes with Applications to Guidance, Interscience Publishers, 1968.

Griffin, R. E., and Sage, A. P., "Large and Small Scale Sensitivity Analysis of Optimum Estimation Algorithms," IEEE Trans. on A.C., Vol. 13, No. 4, August, 1968.

Kushner, H. J., "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes, With Applications," Journal of S.I.A.M. on Control, Ser. A, Vol. 2, No. 1, 1964.

Stratonovich, R. W., "Conditional Markov Processes," Theory and Probability and Its Applications, Vol. 5, No. 2, 1960, pp. 156-78.

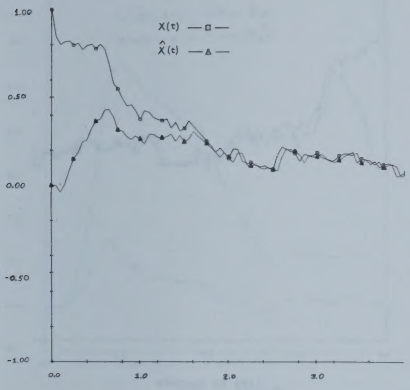


Fig. 1. X(t) and $\hat{X}$(t) vs. time

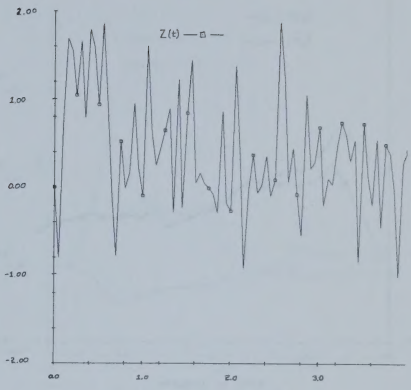


Fig. 2. Z(t) vs. time

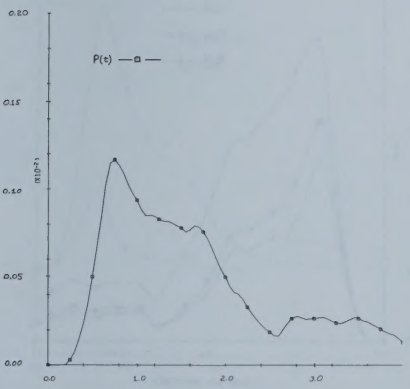


Fig. 3. P(t) vs. time

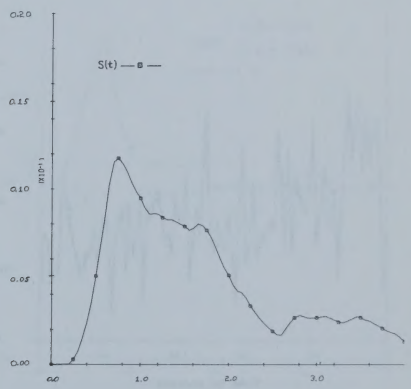


Fig. 4. S(t) vs. time

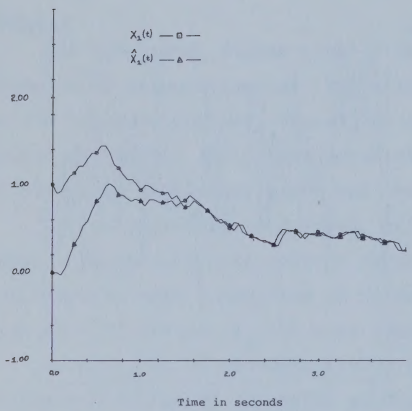


Fig. 5. $X_1(t)$ and $\hat{X}_1(t)$ vs. time

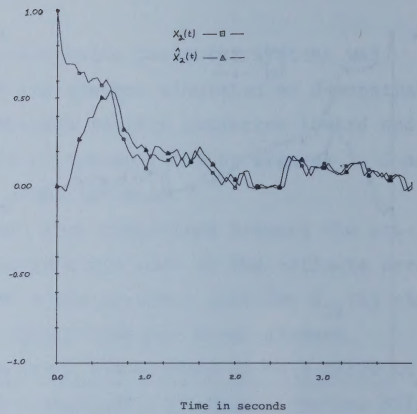


Fig. 6. $X_2(t)$ and $\hat{X}_2(t)$ vs. time

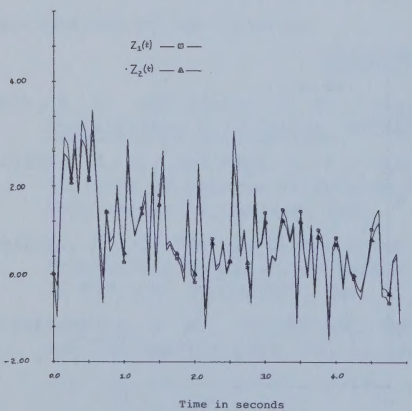


Fig. 7. $Z_1(t)$ and $Z_2(t)$ vs. time

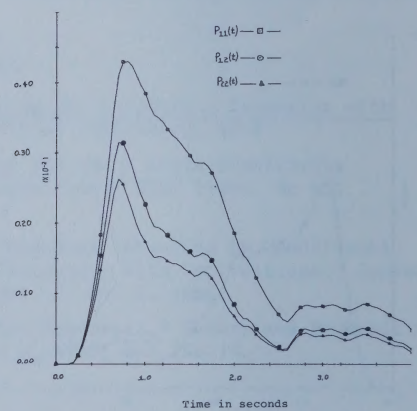


Fig. 8. $P_{11}(t)$, $P_{12}(t)$ and $P_{22}(t)$ vs. time

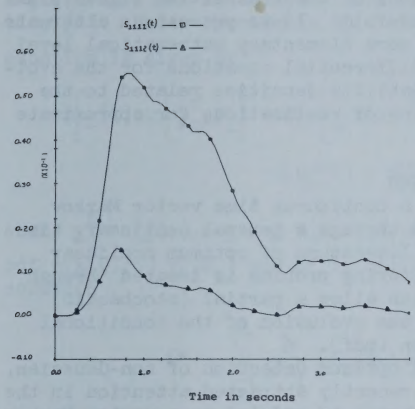


Fig. 9. $S_{1111}(t)$ and $S_{1112}(t)$ vs. time

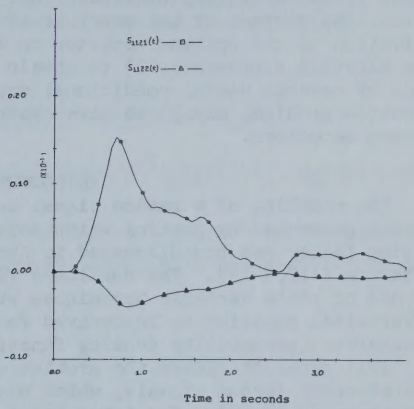


Fig. 10. $S_{1121}(t)$ and $S_{1122}(t)$ vs. time

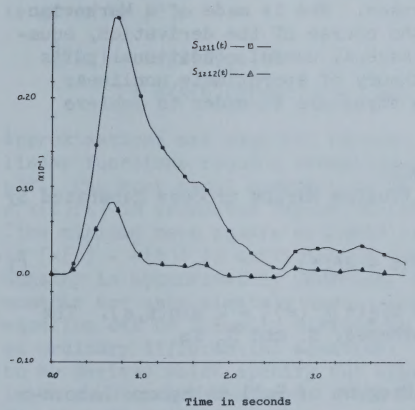


Fig. 11. $S_{1211}(t)$ and $S_{1212}(t)$ vs. time

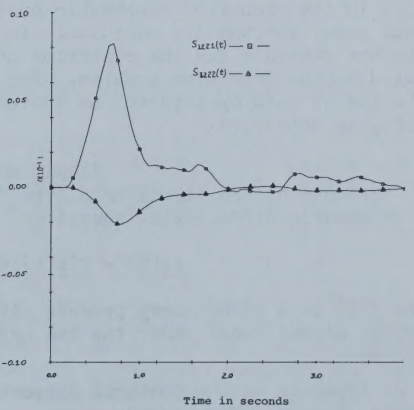


Fig. 12. $S_{1221}(t)$ and $S_{1222}(t)$ vs. time

ON THE DETECTION OF NON-GAUSSIAN SIGNALS

J. J. APPEL* and E. N. PROTONOTARIOS†
Department of Electrical Engineering
Columbia University, New York, N. Y.

ABSTRACT

The optimum detection of non-Gaussian, nonstationary random signals is considered. The transmitted signal is modeled as a vector Markov diffusion process and the received signal is assumed under hypotheses H_1 and H_0 as either of two arbitrary nonlinear functions of the transmitted signal plus noise. The purpose of the paper is three-fold: 1) to present an alternate derivation of the optimum detector on a more elementary mathematical level than existing approaches, 2) to obtain differential equations for the evolution of several useful conditional probability densities related to the detection problem, and 3) to give examples of realizations for approximate optimum detectors.

INTRODUCTION

The modeling of a random signal as a continuous time vector Markov process generated by passing white noise through a general nonlinear, time varying filter has been discussed in the literature of optimum nonlinear filtering [1,2,4-15]. The nonlinear filtering problem is treated through the use of state variable techniques which allow a partial (stochastic) differential equation to be derived for the evolution of the conditional a posteriori probability density function (pdf).

This paper considers the problem of optimum detection of non-Gaussian, nonstationary random signals, which has recently attracted attention in the literature [9,18,19]. The transmitted signal is modeled as a vector Markov diffusion process and the received signal is assumed under hypotheses H_1 and H_0 as either of two arbitrary nonlinear functions of the transmitted signal plus noise having a white component. An alternate derivation of the optimum detector is given here on a more elementary mathematical level than other approaches by Stratonovich-Sosulin [1,2,7], Duncan [17,18], and Kailath [19] which require measure theory, martingales and other advanced topics in the theory of stochastic processes. Use is made of a Markovian signal model and the Itô calculus. In the course of the derivation, equations are obtained for the evolution of several useful conditional pdf's related to the detection problem. The theory of approximate nonlinear filtering is then applied to the optimum structure in order to achieve realizable detectors.

SIGNAL MODEL

Consider as the signal model the diffusion Markov process generated by the stochastic differential equation

$$d\underline{x}(t) = \underline{f}[\underline{x}(t);t]dt + d\underline{\xi}(t) \tag{1}$$

where $\underline{\xi}(t)$ is a Wiener-Levy process with $E[\underline{\xi}(t)\underline{\xi}'(s)] = Q \min(t,s)$. The received signal model under the two hypotheses, H_1 and H_0 is

---

*J. J. Appel is on the Doctoral Support Program of Bell Telephone Laboratories, Holmdel, New Jersey.

†E. N. Protonotarios was supported by the National Science Foundation under Grant GK-4863.

$$d\underline{y}(t) = \begin{cases} \underline{g}[\underline{x}(t);t]dt + d\underline{\eta}(t) & \text{for } H_1 \\ \underline{h}[\underline{x}(t);t]dt + d\underline{\eta}(t) & \text{for } H_0 \end{cases} \tag{2}$$

where $\underline{\eta}(t)$ is a second Wiener-Levy process independent of $\underline{\xi}(t)$, with $E[\underline{\eta}(t)\underline{\eta}'(s)] = N \min(t,s)$.

NONLINEAR FILTERING

If either hypothesis H_1 or H_0 is specified with probability one, the problem becomes one of nonlinear filtering. For this case the conditional a posteriori pdf satisfies a modified Fokker-Planck-Kolmogorov (FPK) equation [4]. Under hypothesis H_1, this partial (stochastic) differential equation is given by

$$dp\left[\underline{x}(t);t \,\middle|\, \underline{y}_0^t, H_1\right] = L^*p\left[\underline{x}(t);t \,\middle|\, \underline{y}_0^t, H_1\right] + p\left[\underline{x}(t);t \,\middle|\, \underline{y}_0^t, H_1\right] \cdot\{d\underline{y}(t) - E\underline{g}[\underline{x}(t);t]dt\}'N^{-1}\{\underline{g}[\underline{x}(t);t] - E\underline{g}[\underline{x}(t);t]\} \tag{3}$$

where $\underline{y}_0^t \equiv \{\underline{y}(\tau), 0 \le \tau \le t\}$; expectations are with respect to the conditional pdf $p[\underline{x}(t);t \mid \underline{y}_0^t, H_1]$ and L^* is the forward operator in the Fokker-Planck equation

$$L^* = -\sum_{i=1}^{m} \frac{\partial}{\partial x_i}\{f_i[\underline{x}(t);t]\} + \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2}{\partial x_i \partial x_j}\{Q_{ij}\} \tag{4}$$

The equation under hypothesis H_0 is identical expect for H_0 replacing H_1, and $\underline{h}[\underline{x}(t);t]$ replacing $\underline{g}[\underline{x}(t);t]$.

A solution to the above stochastic differential equation requires, in general, a solution for an infinite number of moments which completely specify the desired pdf. The stochastic differential equation for the conditional density can be used to derive stochastic differential equations for all these moments. For example, the equation for the conditional mean is given by [5]:

$$d\hat{x}(t) = Ef[x(t);t]dt + E\langle x(t)\{g[x(t);t] - Eg[x(t);t]\}\rangle' N^{-1}\langle dy(t) - Eg[x(t);t]dt\rangle \tag{5}$$

Approximations are required because in general the expectations of the nonlinear functions require knowledge of all the moments of the conditional pdf. The most usual approach is to expand the nonlinearities $f[x(t);t]$ and $g[x(t);t]$ in truncated Taylor series about the current conditional mean [the minimum mean square estimate of $x(t)$] and to assume that the variance of $[x(t) - \hat{x}(t)]$ is sufficiently small that either: (1) the a posteriori density is approximately Gaussian, or (2) the third and higher order central moments are approximately zero. Both sides of the stochastic differential equation can be formally divided by dt and the resulting equation treated as an ordinary differential equation. This allows a pair of coupled equations to be derived which specify the approximate conditional mean and the approximate conditional variance. With the assumption of an approximately Gaussian distribution, the a posteriori pdf is completely specified. These approximate equations have been derived for combinations of the above assumptions. One common form for the scalar case is [6]

$$\dot{\hat{x}} = f(\hat{x}) + \frac{\hat{k}}{N} g'(\hat{x})[r(t) - g(\hat{x})] \tag{6}$$

$$\dot{\hat{k}} = 2f'(\hat{x})\hat{k} - \frac{1}{N}[g'(\hat{x})\hat{k}]^2 + Q + \frac{\hat{k}^2}{N} g''(\hat{x})[r(t) - g(\hat{x})] \tag{7}$$

where $\hat{x}$ is the approximate conditional mean, $\hat{k}$ is the approximate conditional variance and the prime denotes differentiation with respect to $\hat{x}$. The two coupled differential equations for $\hat{x}$ and $\hat{k}$ specify the structure of a nonlinear filter which estimates the conditional mean and thus attempts to provide a minimum mean square estimate for the transmitted signal $x(t)$.

It is interesting to note that the above pair of equations was apparently first derived by Stratonovich from his stochastic differential equation for the conditional a posteriori probability density and was based on his definition of a stochastic integral [7,8] which differs from the more common Itô stochastic integral. Since for both the Itô calculus and the Stratonovich formulation, the resulting partial differential equations for the conditional density functions are equivalent, it is not too surprising that the equations for the first order approximate filters are identical under similar assumptions. The apparent differences between the two formulations can be resolved if one takes into account that the conditional expectation operations appearing in the respective equations must be interpreted with respect to their corresponding stochastic integral.

The Kushner derivation is based on theoretical white noise, so that noise samples are independent even as the interval between samples approaches zero. The Stratonovich derivation is based on "physical" white noise, which means that as the interval between samples is made small enough, a point is reached when noise samples become correlated (this is a result of the fact that all physical systems are band limited).

Both formulations will yield realizable filters assuming that suitable approximations are made. In obtaining a continuous time filter from the Kushner formulation, and assuming the presence of physical white noise instead of theoretical white noise, appropriate additional terms can be introduced [16]. However, the Kushner formulation in finite difference form is directly applicable in the case of digital calculation where the sampling interval is such that the assumption of noise sample independence is valid.

OPTIMUM DETECTION

Consider the Markov process signal model described in Eqs. (1) and (2). Since $\underline{\eta}(t)$ is a vector Wiener-Levy process, if we introduce an increment in time dt, then $d\underline{\eta}(t) \equiv \underline{\eta}(t+dt) - \underline{\eta}(t)$, is a random vector which has a Gaussian distribution with zero mean and covariance matrix Ndt. If $d\underline{y}(t)$ is the incremental change in the observation vector for a dt increment in time, the pdf's for $d\underline{y}(t)$ under hypotheses H_1 and H_0, for $\underline{x}(t)$ known, are Gaussian with mean $\underline{g}[\underline{x}(t);t]$ and covariance matrix Ndt under H_1, and mean $\underline{h}[\underline{x}(t);t]$ and covariance matrix Ndt under H_0.

Consider the joint conditional pdf $p[\underline{x}(t),H_1;t \mid \underline{y}_0^t]$, which is the joint density function for $\underline{x}(t)$ and hypothesis H_1, given the received waveform in the interval $0 \leq \tau \leq t$.

Theorem I

For the signal model given in Eqs. (1) and (2), the joint conditional pdf $p[\underline{x}(t),H_1;t \mid \underline{y}_0^t]$ satisfies the following modified FPK equation

$$dp\left[\underline{x}(t),H_1;t \,\middle|\, \underline{y}_0^t\right] = L^*p\left[\underline{x}(t),H_1;t \,\middle|\, \underline{y}_0^t\right]dt + p\left[\underline{x}(t),H_1;t \,\middle|\, \underline{y}_0^t\right]$$

$$\cdot\{\underline{g}[\underline{x}(t);t] - E[\underline{g},\underline{h}]\}'N^{-1}\{d\underline{y}(t) - E[\underline{g},\underline{h}]dt\} \qquad (8)$$

where

$$E[\underline{g},\underline{h}] \equiv p\left[H_1;t \,\middle|\, \underline{y}_0^t\right]E_1\{\underline{g}[\underline{x}(t);t]\} + p\left[H_0;t \,\middle|\, \underline{y}_0^t\right]E_0\{\underline{h}[\underline{x}(t);t]\} \qquad (9)$$

E_1 denotes expectation with respect to $p[\underline{x}(t);t \mid \underline{y}_0^t,H_1]$ and E_0 denotes expectation with respect to $p[\underline{x}(t);t \mid \underline{y}_0^t,H_0]$.

The proof of this theorem is parallel to the derivation of the modified FPK equation given in Kushner [4] for the nonlinear filtering problem. The difference is that here one considers the joint pdf $p[\underline{x}(t),H_1;t \mid \underline{y}_0^t]$ instead of the a posteriori pdf $p[\underline{x}(t);t \mid \underline{y}_0^t,H_1]$.

From standard statistical decision theory, the optimum decision rule is based on the likelihood ratio, Λ_t, or on a monotonic function of Λ_t, where

$$\Lambda_t = \frac{p\left(\underline{y}_0^t \,\middle|\, H_1\right)}{p\left(\underline{y}_0^t \,\middle|\, H_0\right)} = \frac{p\left(H_1 \,\middle|\, \underline{y}_0^t\right)\pi_0}{\left[1 - p\left(H_1 \,\middle|\, \underline{y}_0^t\right)\right](1-\pi_0)} \qquad (10)$$

where π_0 and $1 - \pi_0$ are the a priori probabilities of hypotheses H_0 and H_1, respectively.

The density function $p[H_1 \mid \underline{y}_0^t]$ will now be determined from the differential equation for the joint conditional density $p[\underline{x}(t),H_1 \mid \underline{y}_0^t]$.

Theorem II

If we assume that the following conditions hold for all i and j,

$$f_i[\underline{x}(t)]p\left(\underline{x}(t) \,\middle|\, H_1,\underline{y}_0^t\right)\Bigg|_{x_i=-\infty}^{x_i=\infty} = 0, \qquad \frac{\partial}{\partial x_j}\left[Q_{ij}p\left(\underline{x}(t) \,\middle|\, H_1,\underline{y}_0^t\right)\right]\Bigg|_{x_i=-\infty}^{x_i=\infty} = 0$$

then $p[H_1;t \mid \underline{y}_0^t]$ satisfies the following stochastic differential equation

$$dp\left[H_1;t \,\middle|\, \underline{y}_0^t\right] = p\left[H_1;t \,\middle|\, \underline{y}_0^t\right]\left\{1 - p\left[H_1;t \,\middle|\, \underline{y}_0^t\right]\right\}$$

$$\cdot\{E_1\underline{g}[\underline{x}(t);t] - E_0\underline{h}[\underline{x}(t);t]\}'N^{-1}\{d\underline{y}(t) - E(\underline{g},\underline{h})dt\} \qquad (11)$$

The proof follows directly by integrating Eq. (8) with respect to $\underline{x}(t)$. Since $p(H_1 \mid \underline{y}_0^t) + p(H_0 \mid \underline{y}_0^t) = 1$ it also follows that $dp(H_0 \mid \underline{y}_0^t) = -dp(H_1 \mid \underline{y}_0^t)$.

At this point we can specify a structure for the optimum detector. It consists of: (1) simulating Eq. (11), subject to the initial condition $p[H_1;t = 0] = 1 - \pi_0$, to obtain $p[H_1;t \mid \underline{y}_0^t]$; (2) evaluating the likelihood ratio in Eq. (10) at the observation time t = T; and (3) comparing $\Lambda(T)$ with a threshold to decide between H_1 and H_0.

It is desirable if a test statistic could be generated directly. Thus we consider the function $Z_t = \ln \Lambda_t$. The differential structure of Z_t is specified by $dZ_t \equiv Z(t+dt) - Z(t)$. Since $p(H_1 \mid \underline{y}_0^t)$ is a solution to stochastic differential equation, we must include terms up to $(dp)^2$ in the expression for the differential of Z_t. In order to obtain the differential of Z_t we make use of the Itô differential rule which states that the differential of a function $f[y(t),t]$, where $y(t)$ is the solution to a stochastic differential equation, is given by

$$df[y,t] = \frac{\partial f[y,t]}{\partial t}\, dt + \frac{\partial f[y,t]}{\partial y}\, dy(t) + \frac{1}{2}\frac{\partial^2 f[y,t]}{\partial y^2}\,[dy(t)]^2 \qquad (12)$$

Theorem III

For the diffusion process signal model in Eqs. (1) and (2), the likelihood ratio Z_t satisfies the following stochastic differential equation

$$dZ_t = \{E_1 g[\underline{x}(t);t] - E_0 h[\underline{x}(t);t]\}' N^{-1} \left[d\underline{y}(t) - \frac{1}{2}\{E_1 g[\underline{x}(t);t] + E_0 h[\underline{x}(t);t]\} dt \right] \qquad (13)$$

subject to the initial condition $Z_0 = 0$.

The proof follows by applying the Itô differential rule in Eq. (12) to the function $Z_t = \ln \Lambda_t$ with $dp[H_1;t \mid \underline{y}_0^t]$ given by Eq. (11).

Kalaith has pointed out that the above structure is completely analogous to the detection of a known signal in white noise (the standard correlation detector) [9]. However, in the case of random signals we must first estimate the signal under each hypothesis and then correlate the received signal with these estimates. Thus, the optimum system for the detection of random signals in noise is an estimator-correlator structure, where the estimator incorporates the optimum nonlinear filter.

The conditional expectations $E_1 g[\underline{x}(t);t]$ and $E_0 h[\underline{x}(t);t]$ imply the presence of two optimum nonlinear filters, since the optimum filters require the complete specification of the conditional density functions $p[\underline{x}(t); t \mid \underline{y}_0^t, H_1]$ and $p[\underline{x}(t);t \mid \underline{y}_0^t, H_0]$ which is also required for the evaluation of the conditional expectations.

Eq. (13) may be formally divided by dt and considered as a differential equation with the observation $\frac{d\underline{y}(t)}{dt} = \underline{r}(t)$ as the input or driving term. The structure of the optimum detector for the case of a non-Gaussian signal in white noise is shown in Fig. 1. The optimum detector is more properly called the optimum receiver, since once the decision H_1 is made, the output of the optimum filter block provides a MMSE estimate of the transmitted signal.

If we let $g[\underline{x}(t);t] \equiv \underline{u}(t)$, the received signal model for the case where H_0 is white noise alone becomes

$$\underline{r}(t) = \begin{cases} \underline{u}(t) + \underline{n}(t) & \text{for } H_1 \\ \underline{n}(t) & \text{for } H_0 \end{cases} \qquad (14)$$

and the optimum detector is

$$\dot{Z}(t) = \hat{\underline{u}}'(t)N^{-1}\underline{r}(t) - \frac{1}{2}\hat{\underline{u}}'(t)N^{-1}\hat{\underline{u}}(t) \qquad (15)$$

$(0,T)$ is the observation interval and $\hat{\underline{u}}(t)$ is the MMSE estimate of $\underline{u}(t) = \underline{g}[\underline{x}(t);t]$ given the received waveform $\{\underline{r}(\tau), 0 \leq \tau \leq t\}$.

Stratonovich and Sosulin [1,2] were the first authors to view the optimum detection problem in terms of optimum nonlinear filtering. Their result for the same case considered above is given by

$$\dot{Z}(t) = \hat{\underline{u}}'(t)N^{-1}\underline{r}(t) - \frac{1}{2}E_1\{\underline{u}'(t)N^{-1}\underline{u}(t)\} \qquad (16)$$

The apparent difference in Eqs. (15) and (16) is due to the two definitions for a stochastic integral. Duncan has shown [17] that in the linear case the Stratonovich result is equivalent to the one based on the Itô calculus if one takes into account the differences in their definitions of a stochastic integral. As we indicated previously the Stratonovich result is valid for the case of physical white noise. This implies that a "real world" continuous time analog detector must be designed using the Stratonovich equation, or else appropriate correction terms must be introduced to the realization based on the Itô calculus. However, the complexity of any optimum nonlinear receiver almost dictates a form of numerical computation. Thus, the finite difference version of the Itô type realization becomes valid. For this reason the detection realizations we will discuss are based on Eq. (13).

APPROXIMATE OPTIMUM DETECTOR

As was pointed out in connection with the optimum nonlinear filter, approximations must be introduced in order to obtain a finite dimensional solution. Consider the case where H_0 is white noise alone. If we assume that the variance of $[\underline{x}(t) - \hat{\underline{x}}(t)]$ is small, then the vector version of Eqs. (6) and (7) specify a nonlinear filter for the approximate conditional mean and conditional variance, $\hat{\underline{x}}$ and $\hat{K}$. If we assume that the a posteriori density is approximately Gaussian, $\hat{\underline{x}}$ and $\hat{K}$ allow us to evaluate $E_1\{\underline{g}[\underline{x}(t);t]\}$. The expression for $E_1\underline{g}[\underline{x}(t);t]$ and Eq. (13) completely specifies the structure of the optimum nonlinear detector.

A second approach is to obtain an approximate optimum nonlinear filter for a MMSE estimate $\hat{\underline{g}}[\underline{x}(t);t] = E_1\{\underline{g}[\underline{x}(t);t]\}$. This structure requires the simulation of three differential equations, one for each of $\hat{\underline{x}}(t)$, $\hat{K}(t)$ and $\hat{\underline{g}}[\underline{x}(t);t]$.

A third approach is to design a best <u>linear</u> filter to estimate $\underline{u}(t)$. We can then make the approximation

$$\tilde{\underline{u}}(t) \equiv \text{best linear estimate of } \underline{u}(t) \approx E_1\underline{g}[\underline{x}(t)\ t]$$

<u>Examples</u>

Consider a random signal spectrum $S_x(\omega) = \dfrac{2a}{\omega^2+a^2}$ and a nonlinearity $g[x(t)] = x(t) + x^3(t)$. The model for this system is given by Eq. (1) with $f[x(t)] = -ax(t)$ and is shown in Fig. 2.

<u>Approximate Detector 1</u> The optimum nonlinear filter based on Eqs. (6) and (7) is specified by

$$\dot{\hat{x}} = -a\hat{x} + \frac{\hat{k}}{N}(1+3\hat{x}^2)[r(t) - (\hat{x}^3+\hat{x})]$$

$$\dot{\hat{k}} = -2a\hat{k} - \frac{1}{N}(1+3\hat{x}^2)^2\hat{k}^2 + 2a + \frac{\hat{k}^2}{N}6\hat{x}[r(t) - (\hat{x}^3+\hat{x})] \qquad (17)$$

and is given in Fig. 3.

If we assume that the conditional density of x(t) is Gaussian with mean, $\hat{x}$, and variance, $\hat{k}$, then

$$\hat{u}(t) = E_1 g[x(t)] = \hat{x}^3 + 3\hat{x}\hat{k} + \hat{x} \tag{18}$$

The structure of the optimum detector is as shown in Fig. 1 with the MMSE filter block consisting of Fig. 3 and Eq. (18).

Approximate Detector 2 Let $h(\cdot)$ be an arbitrary, well behaved scalar function of the state vector $\underline{x}(t)$. Define $\hat{h}(t) = E_1\{h[\underline{x}(t)] \mid \underline{r}_0^t\}$. The MMSE estimate $\hat{h}(t)$ satisfies the equation [4,5]

$$\frac{d}{dt}[\hat{h}(t)] = E_1\{Lh[\underline{x}(t)]\} + \left\{E_1\left[g[\underline{x}(t)]h[\underline{x}(t)]\right] - E_1\left[g[\underline{x}(t)]\right]E_1\left[h[\underline{x}(t)]\right]\right\}' N^{-1} \cdot \left\{\underline{r}(t) - E_1\left[g[\underline{x}(t)]\right]\right\} \tag{19}$$

where L is the backward operator

$$L \equiv \sum_i f_i[\underline{x}(t)] \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_i \sum_j Q_{ij} \frac{\partial^2}{\partial x_i \partial x_j}$$

In the example under consideration $h[\underline{x}(t)] = g[x(t)] = x^3(t) + x(t)$. We make the approximations $E_1\{[g(x)]\} \simeq - a\hat{x}g'(\hat{x}) + \frac{1}{2} Qg''(\hat{x}) = - a\hat{x}(3\hat{x}^2 - 5)$ and $E_1[g^2(x)] - [E_1 g(x)]^2 \approx [g'(x)]^2 \hat{k}$. These approximations are equivalent to keeping the first terms in the appropriate Taylor series expansion around $\hat{x}$. The equation for the approximate MMSE extimate $\hat{g}(t)$ is given by

$$\dot{\hat{g}}(t) = a\hat{x}[5 - 3\hat{x}^2] + \frac{1}{N}[3\hat{x}^2 + 1]^2 \hat{k}[r(t) - \hat{g}(t)] \tag{20}$$

This equation together with Eqs. (17) for $\hat{x}(t)$ and $\hat{k}(t)$, completely specifies the nonlinear filtering block in the optimum detector structure in Fig. 1.

Approximate Detector 3 As a third realization consider the structure employing the best linear filter for estimating $u(t) = g[x(t)]$. For the assumed signal model the spectrum of u(t) is given by

$$S_u(\omega) = \frac{68a\left[\omega^2 + \left(\frac{9}{\sqrt{17}}a\right)^2\right]}{[\omega^2 + (3a)^2][\omega^2 + a^2]} \tag{21}$$

The best linear filter can easily be determined from the theory of Kalman filtering. The approximate optimum detector incorporating this best linear filter is shown in Fig. 4. This structure certainly represents the simpliest realization of the three considered. A comparison of error probabilities for the various approximate realizations will be reported in a future paper.

REFERENCES

[1] R. L. Stratonovich and Y. G. Sosulin, "Optimal Detection of a Markov Process in Noise," Engineering Cybernetics, Vol. 10, No. 6, 1964.

[2] R. L. Stratonovich and Y. G. Sosulin, "Optimum Detection of a Diffusion Process in White Noise," Radio Engineering and Electronic Physics, Vol. 10, No. 5, 1965.

[3] R. L. Stratonovich, "A New Representation for Stochastic Integrals and Equations," SIAM Journal on Control, Vol. 4, No. 2, 1966.
[4] H. J. Kushner, "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes, with Applications," J. SIAM on Control, Vol. 2, No. 1, 1964.
[5] H. J. Kushner, "Dynamical Equations for Optimal Nonlinear Filtering," Journal of Differential Equations, Vol. 3, 1967.
[6] D. L. Snyder, "The State Variable Approach to Analog Communication Theory," IEEE Trans. on IT, Vol. IT-14, No. 1, January 1968.
[7] R. L. Stratonovich, "Conditional Markov Processes," Theory of Probability and Its Applications, Vol. 5, No. 2, 1960.
[8] R. L. Stratonovich, "Application of the Theory of Markov Processes in Optimal Signal Discrimination," Radio Engineering and Electronics, Vol. 5, No. 11, November 1960.
[9] T. Kalaith, 1969 International Symposium on Information Theory, Ellenville, New York, January 28-31, 1969.
[10] F. C. Schweppe, "Evaluation of Likelihood Functions for Gaussian Signals," IEEE Trans. on Info. Theory, Vol. IT-11, No. 1, January 1965.
[11] H. L. Van Trees, "Detection and Continuous Estimation: The Fundamental Role of the Optimum Realizable Linear Filter," WESCON Paper 7.1, August 1966.
[12] R. L. Stratonovich, Conditional Markov Processes and Their Application to the Theory of Optimal Control, American Elsevier, New York, 1968.
[13] R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. ASME J. Basic Engr., Vol. 83, March 1961.
[14] R. S. Bucy, "Nonlinear Filtering Theory," IEEE Trans. Automatic Control (Correspondence), Vol. AC-10, April 1965.
[15] W. M. Wonham, "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering Theory," SIAM J. Control, Vol. 2, 1965.
[16] H. J. Kushner, "Approximations to Optimal Nonlinear Filters," IEEE Trans. of Automatic Control, Vol. AC-12, No. 5, October 1967.
[17] T. E. Duncan, "Probability Densities for Diffusion Processes with Applications to Nonlinear Filtering Theory and Detection Theory," Stanford Electronics Laboratory, May 1967, SEL 67-035, DDC#ZZ14189.
[18] T. Duncan, "Evaluation of Likelihood Functions," Information and Control, Vol. 13, No. 1, July 1968.
[19] T. Kalaith, "A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. on Info. Theory, Vol. IT-15, No. 3, May 1969.

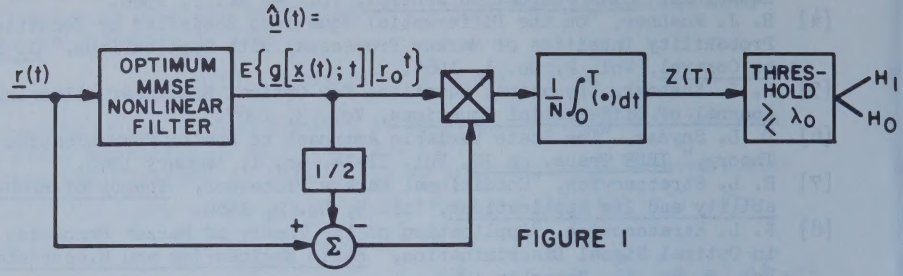


FIGURE 1

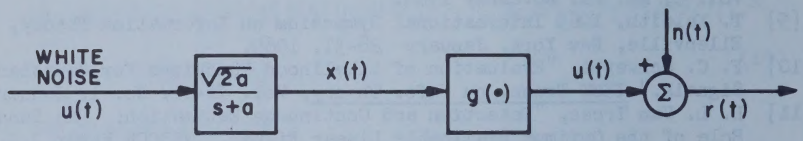


FIGURE 2

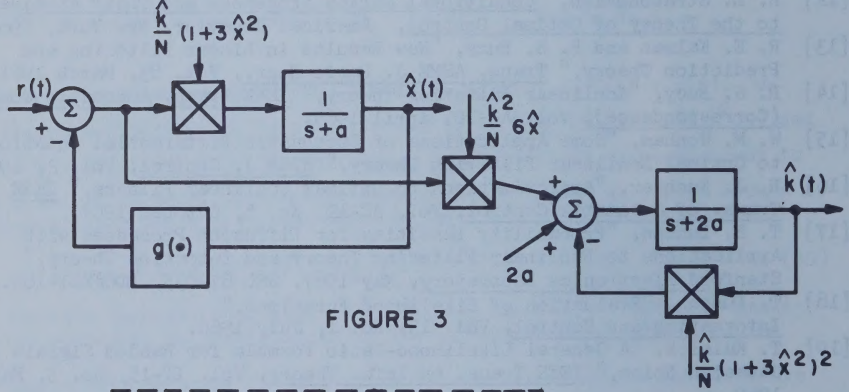


FIGURE 3

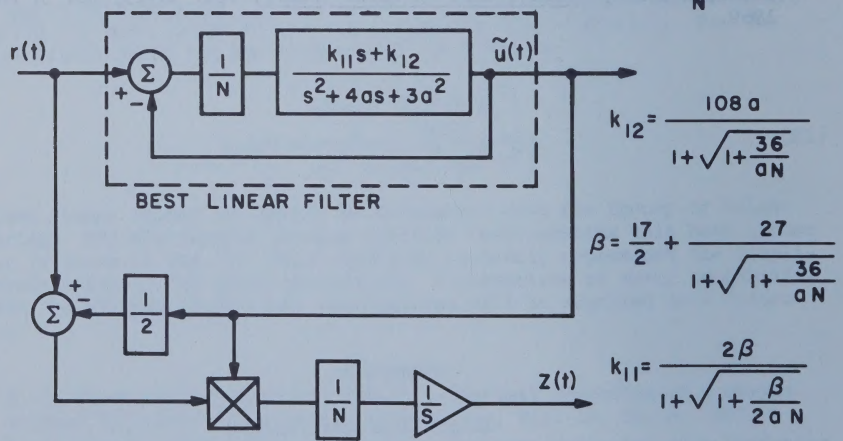


FIGURE 4

PERFORMANCE OF TWO SUBOPTIMUM DETECTORS AND SIGNAL SELECTION IN GAUSSIAN AND IMPULSIVE NOISE

JACK KAPP AND LUDWIK KURZ
Department of Electrical Engineering
New York University
Bronx, New York 10453

ABSTRACT

Two suboptimum, nonlinear, finite memory receivers, selected on the basis of intuitive appeal and ease of implementation, for the noise model represented by a gaussian and impulsive noise mixture, are studied. The performance of these detectors is then compared with that of the optimal in the likehood-ratio-sense detectors. Though the implementation of the proposed detectors is considerably simplified, the performance degradation when compared with the optimum detectors is relatively small. In addition a two-step signal selection procedure which includes bandwidth constraint is considered.

INTRODUCTION

In many real channels both gaussian and impulsive noise are present simultaneously. In this paper, the general philosophy of designing a receiver to operate under large variance noise conditions, with a gaussian noise background which was established in [1,2,3,4] will be extended.

The general philosophy of design is: if one is to operate over a channel in the presence of gaussian noise alone, the receiver output will, under fixed signal conditions, stay in the vicinity of some fixed level, say b, and the optimum detector is a linear filter. However, if the receiver output exceeds this level considerably one intuitively could say that this large deviation can be attributed to the occurrence of an impulse noise and should therefore be discarded.

Using this philosophy, Rappaport and Kurz [2] obtained an optimal decision threshold, Rappaport and Kurz [3] and Kurz [1] obtained an optimal nonlinear device, and Strakhov [4] studied and easily implementable nonlinear device. To simplify analysis, the sample function, $N(t)$, of the channel noise considered in this paper shall be of the following form. (Other noise models may be treated in a similar manner, requiring extensive and sophisticated computer programs.)

$$N(t) = P_o N_1(t) + (1-P_o)N_2(t) \qquad (1)$$

$N_1(t)$ = The Ornstein-Uhlenbeck process, with normalized correlation function $\rho(t_i-t_j) = \exp(-|t_i-t_j|)$

$N_2(t)$ = An impulsive noise process with samples Cauchy distributed with parameter λ, and P_o depends on the severity of the impulsive noise.

It is also assumed that the two noise states occur independently. The closer to one P_o is, the less severe is the impulsive noise. Using a somewhat similar model, Kurz [1] derived an optimal nonlinear device, which is relatively difficult to implement. Using a special case of [1], P_o equal to zero, Rappaport and Kurz [3] obtained an optimal nonlinear device. The second procedure [3] is the one to which the results of this paper are compared.

To overcome the difficulty of implementing the theoretically derived optimal nonlinear devices, two suboptimum detectors are considered. The

---

*The research in this paper was supported by the National Science Foundation under Grant GK-2947. This support is gratefully acknowledged.

first of the two detectors is described mathematically by

$$Z_T = \int_o^T f_I[x(t)]dt \tag{2}$$

where $f_I(x) = -B + (x+B)u(x+B) - (x-B)u(x-B)$ and $x(t)$ is a sample function of the random process representing the input to the detector. The proposed nonlinear device has the advantage of utilizing linear filtering for small signal outputs deviation from the mean, which is an intuitively appealing property. The optimum value of the clipping level, B, for varying impulsive conditions is then determined.

The second nonlinear device is described mathematically by

$$Z_T = \int_o^T f_{II}[x(t)]dt \tag{3}$$

$$f_{II}(x) = -Bu(x+C) + (x+B)u(x+B) - (x-B)u(x-C) - Bu(x-C)$$

where $x(t)$ was defined previously. The nonlinearity $f_{II}(\cdot)$ has the additional feature, when compared with $f_I(\cdot)$ of totally eliminating all large amplitudes exceeding C.

Formulation of the Problem

To overcome the difficulties associated with using a detector based on the likelihood ratio and noting that only the 2nd and 4th semicumulants at the output of the detector are dominant in describing its distribution under "no signal" conditions, a generalized signal-to-noise ratio criterion

$$\frac{S}{N} = \frac{[E(Z_T/x(t) = N(t) + s(t)) - E(Z_T/x(t) = N(t) - s(t))]^2}{Var(Z_T/x(t) = N(t) \pm s(t))} \tag{4}$$

is introduced as a measure of performance of the suboptimum detectors. In (4) Z_T is the output of the detector at t=T (end of decision interval) and antipodal binary signalling with signals $\pm s(t)$ of fixed energy, E_o, and duration, T, is assumed. After the signal-to-noise ratios for the two noise states are calculated, the signal optimization (maximization of (4)) is performed in two stages. In the first stage a signalling class is selected using the criterion

$$C = \max \{E(Z_T/x(t) = N(t) + s(t))\}^2 + \min \{Var(Z_T/x(t) = N(t) + s(t))\} \tag{5}$$

In the second stage, an optimum signal form is selected from the signal class obtained in the first stage. The signal optimization is performed under the assumption of small signal power and large signal duration.

Justification of the Selection of (4) as Optimization Criterion

Considering the gaussian state first, in the no signal case, the moments of the output distribution, which are directly related to the semicumulants, are derived. These moments can be written as

$$m_{2n} = E\{Z^{2n}\} = \int_o^T \cdots \int_o^T R_n(t_1,\ldots,t_n)dt_1 \cdots dt_n \tag{6}$$

where

$$R_n(t_1,\ldots,t_n) = E\{f[x(t_1)] \cdots f[x(t_n)]\} \tag{7}$$

For n=2, using symmetry conditions and integrating by parts, we obtain

$$m_2 = T^2 R_2(0) + 2T \int_{t=o}^{T} (T-t) \frac{\partial R_2(t)}{\partial \rho} \frac{\partial \rho}{\partial t} dt \qquad (8)$$

Using Price's theorem [5]

$$\frac{\partial R_2}{\partial \rho} = E\{ \frac{\partial f[x(t_2)]}{\partial x(t_2)} \frac{\partial f[x(t_2)]}{\partial x(t_2)} \}$$

$$\frac{\partial f}{\partial x} = u(x+B) - u(x-B) \triangleq g(x) \qquad (9)$$

Substituting (9) in (8) and after some mathematical operations, (8) reduces to

$$m_2 = 2T(T+1)\rho(0)\frac{\partial R_2}{\partial \rho}(0) - 2T\rho(T)\frac{\partial R_2}{\partial \rho}(T) + T^2 R_2(0)$$

$$- \frac{2T}{\Pi} \int_{t=o}^{T} (T-1-\sin^{-1}\rho(t))\{EXP(-B^2/(1-\sin t)) - EXP(-B^2/(1+\sin t))\} dt \qquad (10)$$

Derivation of an upper bound for the fourth moment is carried out in a different way. Basically, the nonlinear device, f(x) of (2), is approximated by a Hermite polynomial expansion. The four dimensional Gaussian density function is then represented in terms of its Hermite polynomial expansion. It then becomes a relatively easy, but lengthy process to obtain a bound for the fourth moment.

To find the contribution of the gaussian noise to the output of detector, the results of (6) are extended to four dimensions representation of the Ornstein-Uhlenbeck process. (This has been left out because of the cumbersomeness of the expression.) The nonlinear device is approximated by a Hermite polynomial expansion with the coefficients adjusted to minimize the mean-square error. The four dimensional distribution and the expansion approximation for the nonlinear device are then substituted into (6) and (7) and an upper bound of the fourth moment is obtained. A similar procedure is also used for the impulse noise component.

Since the coefficients of the Gram-Charlier expansion in terms of semi-invariants (6) are related directly to the semi-cumulants and since for gaussian processes only the first two semi-cumulants are nonzero, the representation of the output p.d.f. in this form shows directly the deviation from gaussian distribution (assuming that the edge effects are negligible)as

$$w(Z_T) = \{1 + \frac{1}{4!}\frac{b_4}{\sigma^4} H_4(\frac{Z_T}{\sigma}) + \frac{1}{6!}\frac{b_6}{\sigma^6} H_6(\frac{Z_T}{\sigma}) + \ldots\} w_o(Z_T)$$

where

$$w_o(Z_T) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-1}{2\sigma^2}(Z_T - m_1)\} \qquad (11)$$

Therefore, what must be determined is the upper bound of the second term of (11). This can be achieved by determining the maximum value of this expression for a large observation interval. For instance, for T = 10

$$\max_{B, Z_T} \frac{b_4}{4!\sigma^4} H_4(\frac{Z_T}{\sigma}) = 0.08 \qquad (12)$$

For the gaussian state the output is basically gaussian. Since (12) represents a loose upper bound, it follows that maximum value of the 4th

semi-cumulants will most likely be smaller.

The same type of analysis was carried out for the impulsive state. For instance for T=10, (12) reduced to 0.193 for a range of B in the closed interval [0,3.0]. Thus, for T=10, $B \leq 3$, the output for each noise state has basically been "gaussianized". Since the fourth semi-cumulant is negligible in comparison with the second, the criterion of (4) is meaningful because it depends on the first two semi-cumulants which thus represent the effective contribution of signal and noise to the output.

Selection of the Class of Signals

Before applying (4) it is useful to consider a special case which will give some insight into the type of signals that should be used, i.e., consider the special case of a hard limiter with, b small, and P_o large. In the special case the influence of the impulsive noise component may be neglected. According to [3] when the channel is completely in the impulse noise state, $P_o = 0$, the signal-to-noise ratio used in that paper, a quantity which can be related to (4) is not sensitive to signal shape. Therefore, if there is a signal design problem at all, the channel in its gaussian state will be the state that will have the greatest effect in shaping that signal.

The optimal class of signals is obtained using (5) by maximizing the signal energy and minimizing the noise power. The terms in (5) are obtained by using a Taylor series approximation of the first and second correlation functions about $s(t_1)$ and $s(t_2)$. After removing all the terms independent of $s(t_i)$ and some simplification, (5) reduces to

$$C=\{\phi(\frac{b}{\sqrt{2}})\}^2 \int_{x=o}^{T} \int_{y=o}^{T} s(x)s(y)dxdy - 2 \int_{t=o}^{T} Q_2(t) \int_{x=t}^{T} s(x)s(x-t)dxdt + 2 \int_{t=o}^{T} (-Q_1(t)) \int_{x=t}^{T} s^2(x)dxdt \qquad (13)$$

where $Q_1(t)$ and $Q_2(t)$ are plotted in figure 1. The first integral in (13)is optimized by a square wave. (This can be shown using the calculus of variations.) Based on the form of $Q_1(t)$ and $Q_2(t)$ (Fig. 1) and previous work on on impulsive noise [3,7] it is reasonable to consider the class of signals of the form

$$\begin{aligned} s(t) &= \sqrt{E_o/(T-t_1)}\ \text{Sin}\ \omega t \qquad 0 \leq t \leq t_1 \\ &= \sqrt{E_o/(T-t_1)} \qquad t_1 \leq t \leq T-t_1 \\ &= \sqrt{E_o/(T-t_1)}\ \text{Sin}\ \omega(t-T) \quad T-t_1 \leq t \leq T \end{aligned} \qquad (14)$$

where $\omega = \Pi/2t_1$. (14) has the flexibility of varying the signal shape while keeping the energy constant.

To simplify computer optimization, the second moment of the output is derived differently.

$$R_2(t_1,t_2) = R_2(t_1,t_2)\Big|_{\rho(\tau)=0} + \int_0^{\rho(\tau)} \frac{\partial R_2}{\partial \rho} d\rho \qquad (15)$$

$$\frac{\partial R_2}{\partial \rho} = \frac{\partial R_2}{\partial \rho}\Big|_{\rho(\tau)=0} + \int_0^{\rho(\tau)} \frac{\partial^2 R_2}{\partial \rho^2} d\rho$$

The major difference in this derivation is that in the expectation operation of (15) the two dimensional Hermite polynomial expansion of the gaussian process was used, thus obtaining

$$m_2 = \{\phi(\frac{B}{\sqrt{2}})\}^2 \int_o^T \int_o^T \rho(t_1,t_2)dt_2dt_2$$

$$+ 4 \sum_{n=o}^{+\infty} \frac{\{w(B)H_{2n-1}(B)\}^2}{(2n+3)!} \int_o^T \int_o^T [\rho(t_1,t_2)]^{2n+3}\, dt_1dt_2 \qquad (16)$$

Equation (16) though cumbersome converges rapidly.

For the small signal case, a Taylor series expansion of (15) with respect to $s(t_1)$,$s(t_2)$, can be taken. The signal energy is lowered until only the first three terms of (15) affect m_2.

Figure 2 shows S/N as a function of t_1. At every t_1 the signal-to-noise was taken at the value of B which optimized it for that value of t_1. From Figure 2 it can be observed that the larger the bandwidth (the smaller t_2) the greater the S/N ratio.

Signal Design

For a large value of P_o (see Figure 2), the signal-to-noise ratio is sensitive to signal shape; this will be exploited in the signal design problem. The problem reduces to finding an appropriate signal in the class of (14), or equivalently, to finding an optimum t_1. It is useful to limit the frequency range, Δf, of the signal. The criterion that was used is

$$C = w \max_{t_1}(s/n) + \min_{t_1}(\Delta f) \qquad (17)$$

where w is a weighting factor that is selected from physical and economical considerations (relative cost of S/N and bandwidth). Some values of the optimal t_1 for $P_o = 0.99$ are given in the table below.

w	t_1 optimal
5	0.7
4	0.8
3	0.95

As P_o decreases t_1 increases and the frequency term becomes the relevant term in (17); or the S/N is less sensitive to t_1.

Second Processing Procedure

The second processing procedure investigated is described mathematically by (3). This device has the additional advantage of completely eliminating any impulsive noise sample that exceeds C.

After testing the semi-cumulants of the p.d.f. of the output distribution, it can be shown that the S/N ratio is a good criterion of performance; in fact, in the impulsive state, for finite C, the second term of (11) is smaller than for the first processing procedure.

Using a similar procedure as before, the equations of the S/N ratio were obtained. Figure 3 shows plots of the S/N vs B, with t_1 as parameter. For every B an optimum value of C was used. The dotted graphs represent the S/N as a function of B for the first processing procedure.

The reason for the improved performance of detector (3) when compared to detector (2) can be partially traced to the choice of the impulsive noise model. The model used represents an extremely severe noise condition: totally eliminating the large samples is far better than simply clipping them. In addition, in the detector (3) a sharp maximum appears (see Fig.3). This results in an increased sensitivity of the detector to the channel conditions. This undesirable condition does not appear in the detector (2).

In Figure 4 S/N ratio is shown as a function of the frequency range. Again for detector (3), the sensitivity to channel conditions can be observed. Another fact that is of interest is the plotting out of the S/N ratio as P_o decreases. Which indicates that for small P_o the signal shape does not affect the performance of the system. For P_o greater than.5 the S/N is very sensitive to signal shape.

A signal design using (17) was carried out. Some results are listed below.

P_o	w	t_1
0.99	5	0.021
0.99	4	0.043
0.99	3	0.076

Comparison to Other Work

As P_o becomes zero the channel is in a completely impulsive state. These channel conditions were studied in [3] and an optimal nonlinear detector was derived. The detector performance was the signal-to-noise ratio of the output random variable. The criterion used, which reduced the likelihood ratio to a gaussian problem, can be compared to the criterion of this paper on the basis of the effective S/N ratio, namely,

$$\frac{S}{N} = 4.T\rho^2 \tag{18}$$

where ρ is the signal-to-noise ratio used in [3]. The two detectors, the optimal of [3] and the detector of (3) were compared for a square wave, and came to 95.627% of the optimal system with $\lambda = 5$ and $E_o = 1.25$, which indicates that the detector of (3) compares favorably with the optimal detector of [3] when the channel is in its noise state, $P_o = 0$.

Estimation of P_o

As mentioned before, the optimal value of B,C, and t_1 are sensitive to P_o which controls the influence of the impulsive noise on the system optimization. In practical systems the parameter P_o is not known and varies. A procedure for estimating and updating P_o was given in [8]. The subroutine of [8] can be integrated into the optimization computer program developed for the present system. Thus, obtaining a system which adopts to the changing channel conditions.

REFERENCES

1. L. Kurz, "A Method of Digital Signalling in the Presence of Additive Gaussian and Impulsive Noise", 1962 IRE International Convention Record, Part 4, Vol. 10, pp. 161-173.

2. S.S. Rappaport and L. Kurz, "Optimal Decision Thresholds for Digital Signalling in Non-Gaussian Noise", 1965 IEEE International Convention Record, Part 2, Vol. 13, pp. 198-212.

3. S.S. Rappaport and L. Kurz, "An Optimal Nonlinear Detector for Digital Data Transmission Through Non-Gaussian Channels", IEEE Transactions on Communication Technology, Vol. Com-14, June 1966, pp. 266-274.

4. Nicholas A. Strakhov, "A Representation Theory for a Class of Non-Gaussian Distributions with Applications to Digital Data Transmission", New York University Thesis, October 1967.

5. R. Price, "A Useful Theorem for Nonlinear Devices Having Gaussian Inputs", IRE Transactions on Information Theory, Vol. IT-4, June 1968, pp. 59-72.

6. R.L. Stratonovich, Topics in the Theory of Random Noise, Volume I., Gordon and Breach.

7. H. Silver and L. Kurz, "A Signal Design Problem in Burst Noise Using Game Techniques," Proceedings of the Second Annual Princeton Conference on Information Theory and Systems, pp. 114.

8. S. Cohen and L. Kurz, "An Efficient Nonparametric Detection Procedure Using Computer Selection Techniques," Presented to the Polytechnic Institute of Brooklyn Symposium on Computer Processing in Communications, April 10, 1969, Vol. XIX of the MRI Symposium Series.

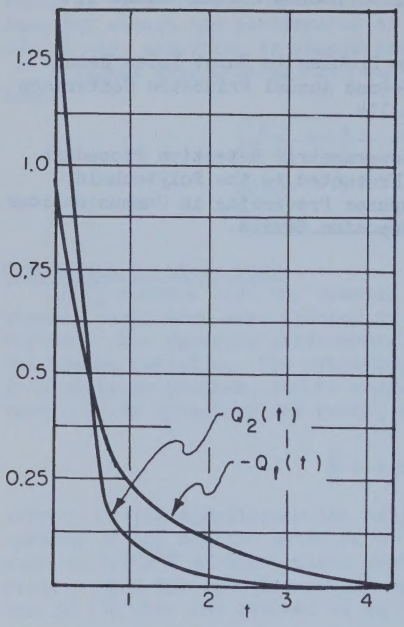


FIG. 1

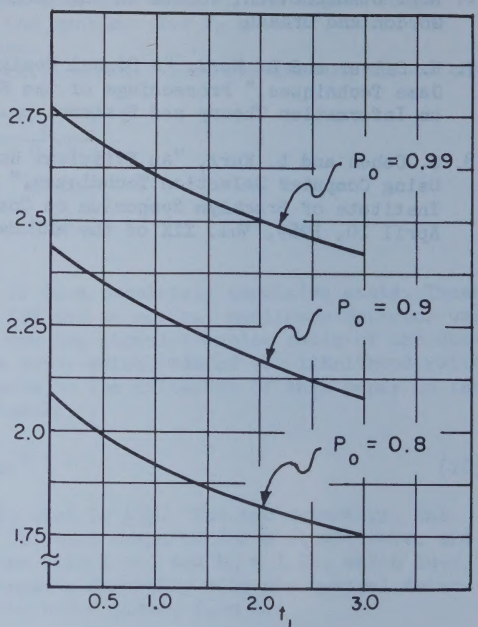


FIG. 2

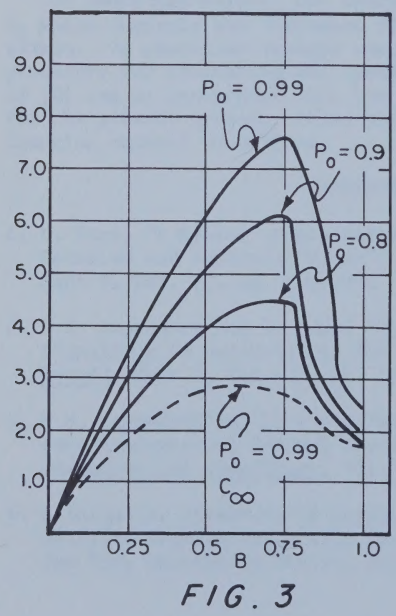


FIG. 3

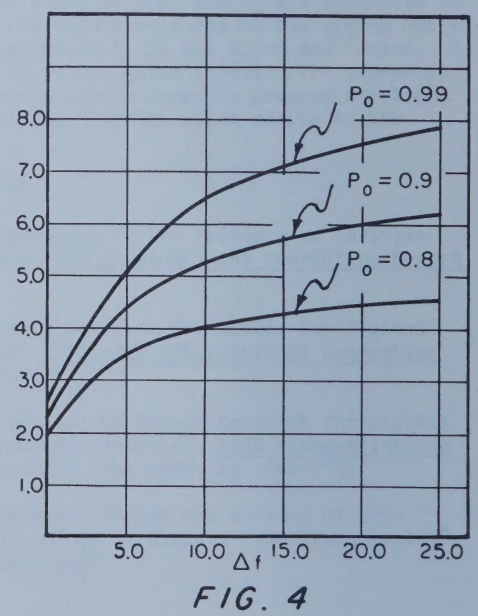


FIG. 4

SEQUENTIAL ESTIMATION FOR NONLINEAR SYSTEMS WITH GAUSS-MARKOV OBSERVATION NOISE

N. K. LOH
Department of Electrical Engineering
The University of Iowa
Iowa City, Iowa 52240

ABSTRACT

The problem of multichannel discrete-time nonlinear filtering and smoothing with Gauss-Markov observation noise is investigated. The approach is based on the method of maximum likelihood. Kalman's formal limiting procedure is applied to the sequential discrete-time estimators to obtain the corresponding continuous-time filter and smoother. The results obtained here reduce to known results when the system is linear.

I. INTRODUCTION

The theory of linear filtering with additive Gaussian white-noise disturbances is well known [1]; the generalization to the case where observations contain Gauss-Markov noise is recognized as an important problem and was considered in [2]-[7]. The problem of multichannel nonlinear filtering and smoothing with Gauss-Markov observation noise is investigated in this paper. The results obtained here reduce to known results [2]-[7], when the system is linear.

Sections III and IV treat the discrete-time sequential estimation problems. In section V, Kalman's formal limiting procedure [8] is applied to obtain the corresponding continuous-time filter and smoother. The continuous-time filter obtained in this paper agree with the results obtained in [9], where a different method is employed. The developement here is based on the method of maximum likelihood [10]-[12]. We emphasize that the derivations are formal.

II. PROBLEM STATEMENT AND FORMULATION

The dynamical system under consideration is described by

$$x_{k+1} = f(x_k,k)+G_k\xi_k, \quad (x: nx1,\ \xi: rx1)$$
$$x(o) = x_o, \quad x_o \sim N[\bar{x}_o, X_o], \tag{1}$$

with observations

$$y_{k+1} = H_{k+1}x_{k+1}+\nu_{k+1}, \quad (y: mx1) \tag{2}$$

where the Gauss-Markov noise ν_k satisfies

$$\nu_{k+1} = B_k\nu_k + F_k\eta_k, \quad (\nu: \text{mx1},\ \eta: \text{sx1})$$
$$\nu(o) = \nu_o, \quad \nu_o \sim N[\bar{\nu}_o, V_o]. \tag{3}$$

Here $\{\xi_k, k=0,1,\ldots\}$ and $\{\eta_k, k=0,1,\ldots\}$ are zero-mean Gaussian white-noise sequences with $E[\xi_j\xi_k^T] = Q_k\delta_{jk}$, $E[\eta_j\eta_k^T] = N_k\delta_{jk}$ and $E[\xi_j\eta_k^T] = C_k\delta_{jk}$, where E denotes expectation, and δ_{jk} is the Kronecker delta. Further ξ_k and η_k are assumed to be uncorrelated with x_o and ν_o for all $k \geq 0$.

Since the filtering problem described by Eqs. (1)-(3) contains Gauss-Markov noise, we wish to put it into a standard filtering problem with observations containing only Gaussian white-noise sequences. To accomplish this, we define

$$z_k \triangleq y_{k+1} - B_k y_k = h(x_k,k) + \alpha_k, \tag{4}$$

where

$$h(x_k,k) = H_{k+1}f(x_k,k) - B_kH_kx_k \tag{5}$$

$$\alpha_k = H_{k+1}G_k\xi_k + F_k\eta_k \tag{6}$$

$$E[\alpha_j\alpha_k^T] = R_k\delta_{jk}$$

$$R_k = H_{k+1}G_kQ_kG_k^TH_{k+1}^T + F_kN_kF_k^T + H_{k+1}G_kC_kF_k^T + F_kC_k^TG_k^TH_{k+1}^T, \tag{7}$$

and R_k is assumed to be positive-definite for all $k \geq 0$. The sequence of 'derived' observations $Z_k = \{z_k, k=0,1,\ldots\}$ now contains only Gaussian white-noise sequences. Note that ξ_k and α_k are correlated. This method of 'whitening' the observations is due to Bryson and Henrikson [6].

Utilizing Eqs. (1) and (4), the problem now is to obtain a maximum likelihood estimate $\hat{x}_{k|j}$ of x_k, given the sequence of observations $Z_j = \{z_j, j=0,1,\ldots\}$. For k=j, $x_{k|k}$ is a filtered estimate; for k>j, $x_{k|j}$ is a predicted estimate; and for j=N > k, $\hat{x}_{k|N}$ is a smoothed estimate. However, in view of the fact that z_k is calculated from y_{k+1}, the filtered estimate $\hat{x}_{k|k}$ given Z_k is actually the single-stage smoothed estimate $\hat{x}(k|Y_{k+1})$ given $Y_{k+1} = \{y_o,\ldots,y_{k+1}\}$. Similarly, the single-stage predicted estimate $\hat{x}_{k|k-1}$ given Z_{k-1}, and the smoothed estimated $\hat{x}_{k|N}$ given Z_N, are actually the filtered estimate $\hat{x}(k|Y_k)$ given Y_k and the smoothed estimate $\hat{x}(k|Y_{N+1})$ given Y_{N+1}, respectively. Here, we shall consider only the estimate $\hat{x}_{k|j}$ based on Z_j.

We consider the likelihood functions [10]-[12]:

$L(x_k,z_k) \triangleq \ln p(x_k|Z_k)$ for filtering;

$L(x_k,Z_{k-1}) \triangleq \ln p(x_k|Z_{k-1})$ for predicting; and

$L(x_k,x_{k+1},Z_N) \triangleq \ln p(x_k,x_{k+1}|Z_N)$ for smoothing (N > k), where $p(x_k|Z_j)$ is the

conditional density function of x_k given Z_j. The maximum likelihood estimates of x_k are the values of x_k which maximize the above likelihood functions.

III. SOLUTIONS OF THE PREDICTING AND FILTERING PROBLEMS

We wish to maximize, with respect to x_{k+1}, the likelihood function,

$$L(x_{k+1},Z_{k+1}) \triangleq \ell n\ p(x_{k+1}|Z_{k+1})$$
$$= \ell n[p(z_{k+1}|x_{k+1})p(x_{k+1}|Z_k)/p(z_k|Z_{k-1})] \tag{8}$$

with $p(z_{k+1}|x_{k+1})$ known, i.e.,

$$p(z_{k+1}|x_{k+1}) = c_1 \exp\{-\tfrac{1}{2}||z_{k+1}-h(x_{k+1},k+1)||^2_{R^{-1}_{k+1}}\}, \tag{9}$$

where $||a||^2_A = a^T A a$, and c_1 is a normalization constant (all subsequent c_i are normalization constants). Since $p(z_k|Z_{k-1})$ does not depend on x_{k+1}, we only have to compute the predicting density $p(x_{k+1}|Z_k)$ in Eq. (8).

The predicting density $p(x_{k+1}|Z_k)$ can be expressed as

$$p(x_{k+1}|Z_k) = \int_{-\infty}^{\infty} p(x_{k+1}|x_k,Z_k)p(x_k|Z_k)dx_k. \tag{10}$$

Now by the conditional expectation theorem on Gaussian random variables (see, for example, [13]), we have

$$p(x_{k+1}|x_k,Z_k) = p(x_{k+1} = f(x_k,k)+G_k\xi_k|x_k,Z_{k-1},z_k = h(x_k,k)+\alpha_k)$$
$$= c_2 \exp\{-\tfrac{1}{2}||x_{k+1}-f(x_k,k)-S_k[z_k-h(x_k,k)]||^2_{\bar{Q}^{-1}_k}\}, \tag{11}$$

where

$$S_k = (G_k Q_k G_k^T H_{k+1}^T + G_k C_k F_k^T) R_k^{-1} \tag{12}$$

$$\bar{Q}_k = G_k Q_k G_k^T - S_k R_k S_k^T . \tag{13}$$

The integration in Eq. (10) can be carried out in closed form if we make the following approximations,

$$p(x_k|Z_k) = c_3 \exp\{-\tfrac{1}{2}(x_k-\hat{x}_k|_k)^T P^{-1}_{k|k}(x_k-\hat{x}_k|_k)\}$$
$$f(x_k,k) = f(\hat{x}_{k|k},k)+f_x(\hat{x}_{k|k},k)(x_k-\hat{x}_{k|k}) \tag{14}$$
$$h(x_k,k) = h(\hat{x}_{k|k},k)+h_x(\hat{x}_{k|k},k)(x_k-\hat{x}_{k|k}) ,$$

where $\hat{x}_{k|k}$ and $P_{k|k}$ have yet to be determined, $f_x(\hat{x}_{k|k},k)$ and $h_x(\hat{x}_{k|k},k)$ are, respectively, the nxn and mxn Jacobian matrices of $f(x_k,k)$ and $h(x_k,k)$ evaluated at $x = \hat{x}_{k|k}$.

Substituting Eqs. (11) and (14) into (10) and performing the integration, we obtain

$$p(x_{k+1}|Z_k) = c_4 \exp\{-\tfrac{1}{2}(x_{k+1}-\hat{x}_{k+1|k})^T P^{-1}_{k+1|k}(x_{k+1}-\hat{x}_{k+1|k})\}, \tag{15}$$

where

$$\hat{x}_{k+1|k} \triangleq f(\hat{x}_{k|k},k)+S_k[z_k-h(\hat{x}_{k|k},k)] \tag{16}$$

$$P_{k+1|k} \triangleq [f_x(\hat{x}_{k|k},k)-S_k h_x(\hat{x}_{k|k},k)]P_{k|k}[f_x(\hat{x}_{k|k},k)-S_k h_x(\hat{x}_{k|k},k)]^T+\bar{Q}_k. \tag{17}$$

Obviously, $p(x_{k+1}|Z_k)$ is normal with conditional mean $\hat{x}_{k+1|k}$ and conditional covariance $P_{k+1|k}$ given by Eqs. (16) and (17), respectively.

Now from Eqs. (8), (9) and (15), we see that maximization of $L(x_{k+1}, Z_{k+1})$ is equivalent to the minimization of a performance criterion of the form

$$J_1 = ||z_{k+1}-h(x_{k+1},k+1)||^2_{R^{-1}_{k+1}} + ||x_{k+1}-\hat{x}_{k+1|k}||^2_{P^{-1}_{k+1|k}}. \tag{18}$$

A necessary and sufficient condition for determining the maximum likelihood filtered estimate $\hat{x}_{k+1|k+1}$ is

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k}+[I-P_{k+1|k}a_x(\hat{x}_{k+1|k},z_{k+1})]^{-1}P_{k+1|k}h^T_x(\hat{x}_{k+1|k},k+1)R^{-1}_{k+1}[z_{k+1}-h(\hat{x}_{k+1|k},k+1)] \tag{19}$$

where we have made use of the approximation

$$a(x_{k+1},z_{k+1}) \triangleq h^T_x(x_{k+1},k+1)R^{-1}_{k+1}[z_{k+1}-h(x_{k+1},k+1)]$$
$$= a(\hat{x}_{k+1|k},z_{k+1})+a_x(\hat{x}_{k+1|k},z_{k+1})(x_{k+1}-\hat{x}_{k+1|k}), \tag{20}$$

and $a_x(\hat{x}_{k+1|k},k+1)$ is the mxn Jacobian matrix of $a(x_{k+1},k+1)$ evaluated at $x = \hat{x}_{k+1|k}$.

The associated filter error covariance may be shown to be approximately given by

$$P_{k+1|k+1} \triangleq E[(x_{k+1}-\hat{x}_{k+1|k+1})(x_{k+1}-\hat{x}_{k+1|k+1})^T|Z_{k+1}]$$
$$= [I-K_{k+1|k}h_x(\hat{x}_{k+1|k},k+1)]P_{k+1|k}[I-K_{k+1|k}h_x(\hat{x}_{k+1|k},k+1)]^T+$$
$$+K_{k+1|k}R_{k+1}K^T_{k+1|k} \tag{21}$$

$$K_{k+1|k} = [I-P_{k+1|k}a_x(\hat{x}_{k+1|k},z_{k+1})]^{-1}P_{k+1|k}h^T_x(\hat{x}_{k+1|k},k+1)R^{-1}_{k+1}.$$

The initial conditions for $\hat{x}_{k+1|k+1}$ and $P_{k+1|k+1}$ are determined by computing $p(x_o|Z_o) \equiv p(x_o|z_o)$,

$$p(x_o|z_o) = p(z_o|x_o)p(x_o)/\int_{-\infty}^{\infty} p(z_o|x_o)p(x_o)dx_o = c_5 \exp\{-\tfrac{1}{2}||x_o-\hat{x}_o||^2_{P_o^{-1}}\}, \quad (22)$$

where

$$\hat{x}_o = \bar{x}_o + X_o h_x^T(\bar{x}_o,t_o)[h_x(\bar{x}_o,t_o)\,X_o h_x^T(\bar{x}_o,t_o)+R_o]^{-1}[z_o-h(\bar{x}_o,t_o)]$$

$$P_o = X_o - X_o h_x^T(\bar{x}_o,t_o)[h_x(\bar{x}_o,t_o)X_o h_x^T(\bar{x}_o,t_o)+R_o]^{-1}h_x(\bar{x}_o,t_o)X_o. \quad (23)$$

Equations (19) and (21) together with Eqs. (16), (17) and (23) provide the required sequential computational algorithm for the maximum likelihood estimate $\hat{x}_{k+1|k+1}$. Recall that $\hat{x}_{k+1|k+1}$ is actually a single-stage smoothed estimate. Similarly, the filter error covariance $P_{k+1|k+1}$ is a single-stage smoother error covariance. In order to compute $\hat{x}_{k+1|k+1}$ and $P_{k+1|k+1}$, we have to take $Y_{k+2} = \{y_o,\ldots,y_{k+2}\}$ observations, i.e., one observation ahead of an estimate. When the system is linear, the results obtained here reduce to those of references [6] and [7], and the single-stage smoother may be rearranged to become a filter [6].

IV. SOLUTION OF THE SMOOTHING PROBLEM

We consider the likelihood function

$$L(x_k,x_{k+1},Z_N) \triangleq \ell n\, p(x_k,x_{k+1}|Z_N)$$

$$Z_N = \ell n[p(x_{k+1}|x_k,Z_k)p(x_k|Z_k)p(z_{k+1},\ldots,z_N|x_{k+1})p(Z_k)/p(Z_N)] \quad (24)$$

Proceeding as in Section III, it follows that we have to maximize

$$\max_{x_k,x_{k+1}} L(x_k,x_{k+1},Z_N) = \max_{x_k,x_{k+1}} \{-\tfrac{1}{2}||x_{k+1}-f(x_k,k)-S_k[z_k-h(x_k,k)]||^2_{Q_k^{-1}} - \tfrac{1}{2}||x_k-x_{k|k}||^2_{P_{k|k}^{-1}} + J\} \quad (25)$$

where J contains all terms, including $\ell n\, p(z_{k+1},\ldots,z_N|x_{k+1})$, which do not involve x_k. It follows immediately that $\hat{x}_{k|N}$ is that value of x_k which minimizes

$$J_2 = ||\hat{x}_{k+1|N}-f(x_k,k)-S_k[z_k-h(x_k,k)]||^2_{Q_k^{-1}} + ||x_k-\hat{x}_{k|k}||^2_{P_{k\,k}^{-1}}. \quad (26)$$

A necessary and sufficient condition for determining the sequential smoothed estimate $\hat{x}_{k|N}$ is

$$\hat{x}_{k|N} = \hat{x}_{k|k} + A_{k|N}[\hat{x}_{k+1|N}-\hat{x}_{k+1|k}], \quad (27)$$

where

$$A_{k|N} = [I-P_{k|k}b_x(\hat{x}_{k|k},\hat{x}_{k+1|N},z_k)]^{-1}P_{k|k}[f_x(\hat{x}_{k|k},k)-S_k h_x(\hat{x}_{k|k},k)]^T Q_k^{-1} \quad (28)$$

$$b(x_k,\hat{x}_{k+1|N},z_k) \triangleq [f_x(x_k,k)-S_k h_x(x_k,k)]^T Q_k^{-1}\{\hat{x}_{k+1|N}-f(x_k,k)-S_k[z_k-h(x_k,k)]\}$$
$$= b(\hat{x}_{k|k},\hat{x}_{k+1|N},z_k)+b_x(\hat{x}_{k|k},\hat{x}_{k+1|N},z_k)(x_k-\hat{x}_{k|k}), \quad (29)$$

and $b_x(\hat{x}_{k|k},\hat{x}_{k+1|N},z_k)$ is the nxn Jacobian matrix of $b(x_k,\hat{x}_{k+1|N},z_k)$ evaluated at $x = \hat{x}_{k|k}$. The filtered estimate $\hat{x}_{k|k}$ and the predicted estimate $\hat{x}_{k+1|k}$ are as given in Eqs. (19) and (16), respectively. The computation of $\hat{x}_{k|N}$ is initiated by specifying $\hat{x}_{N-1|N-1}$, $\hat{x}_{N|N}$, and $\hat{x}_{N|N-1}$ when k = N-1. The filtered estimate $\hat{x}_{k|k}$ must be stored for all k = N-1,N-2,...,1,0.

The sequential relation for the smoother error covariance may be shown to be approximately given by

$$P_{k|N} \triangleq E[x_k-\hat{x}_{k|N})(x_k-\hat{x}_{k|N})^T|Z_N]$$
$$= P_{k|k}-A_{k|N}[P_{k+1|k}-P_{k+1|N}]A_{k|N}^T \quad (30)$$

with the initial conditions $P_{N-1|N-1}$, $P_{N|N-1}$, and $P_{N|N}$ when k = N-1. The filter error covariance and the predictor error covariance $P_{k+1|k}$ are as given in Eq. (21) and (17), respectively. The filter error covariance $P_{k|k}$ must be stored for all k = N-1,N-2,...,1,0.

As pointed out in Section II, the smoothed solutions $\hat{x}_{k|N}$ and $P_{k|N}$ conditioned on Z_N here are actually the smoothed solutions $\hat{x}(k|Y_{N+1})$ and $P(k|Y_{N+1})$ conditioned on Y_{N+1}. As expected, Eqs. (27) and (30) reduce to known results for the special case of linear smoothing [6], [7].

V. THE CONTINUOUS-TIME FILTER AND SMOOTHER

The continuous-time analogue of the discrete-time estimation problem is described by the following system of stochastic differential equation*

$$\dot{x}(t) = f(x(t),t) + F(t)\xi(t)$$
$$x(t_o) = x_o,\ x_o \sim N[\bar{x}_o,X_o], \quad (31)$$
$$y(t) = H(t)x(t) + \nu(t), \quad (32)$$
$$\dot{\nu}(t) = B(t)\nu(t) + F(t)\eta(t)$$
$$\nu(t_o) = \nu_o,\ \nu_o \sim N[\bar{\nu}_o,V_o], \quad (33)$$

where the dimensions of the various vectors and matrices remain the same as in Eqs. (1)-(3); $\xi(t)$ and $\eta(t)$ are zero-mean white-noise processes with $E[\xi(t)\xi^T(\tau)] = Q(t)\delta(t-\tau)$, $E[\eta(t)\eta^T(\tau)] = N(t)\delta(t-\tau)$ and $E[\xi(t)\eta^T(\tau)] = C(t)\delta\ t-\tau)$; $x(t_o)$ and $\nu(t_o)$ are not correlated with each other and are also not correlated with $\xi(t)$ and $\eta(t)$ for all $t \geq 0$.

---

* Equation (31) is to be interpreted in the sense of Ito [14]-[15], and is normally written as $dx(t) = f(x(t),t)dt+G(t)dw(t)$, where w(t) is a vector Wiener process.

We note that by applying Ito's Lemma [16] to y(t) in conjunction with Eqs. (31) and (33), we have [9]

$$\dot{y}(t)-B(t)y(t) = h(x(t),t)+\alpha(t), \tag{34}$$

where

$$h(x(t),t) = H(t)f(x(t),t)+[\dot{H}(t)-B(t)H(t)]x(t) \tag{35}$$

$$\alpha(t) = H(t)G(t)\xi(t)+F(t)\eta(t). \tag{36}$$

The right-hand side of Eq. (34) is the continuous-time analogue z(t) of the derived observation z(k). However, we shall obtain the relation in Eq. (34) by formally applying the limiting procedure of Kalman [8] to Eq. (4). The various analogues between the discrete-time and continuous-time models are needed. These are given in the Appendix (see also [8] and [12]).

With the help of the Appendix, we readily obtain, to within the first-order terms in Δt,

$$\begin{aligned} z(t) &\triangleq y(t+\Delta t)-[I+B(t)\Delta t]y(t) \\ &= [\dot{y}(t)-B(t)y(t)]\Delta t \end{aligned} \tag{37}$$

where $[y(t+\Delta t)-y(t)]/\Delta t \to \dot{y}(t)$ as $\Delta t \to 0$. Hence z(k) in the discrete-time model is to be replaced by $[\dot{y}(t)-B(t)y(t)]\Delta t$ in the continuous-time model. Other continuous-time analogues may be established in the same way.

(a) Continuous-time Filter Equations

We have the following analogues to within the first-order terms in Δt:

$$\begin{aligned} \hat{x}_{k+1|k} &\equiv \hat{x}(t+\Delta t|t) \\ &= \hat{x}(t|t)+f(\hat{x}(t|t),t)\Delta t+S(t)[\dot{y}(t)-B(t)y(t)-h(\hat{x}(t|t),t)]\Delta t \end{aligned}$$

$$\begin{aligned} P_{k+1|k} &\equiv P(t+\Delta t|t) \\ &= P(t|t)+\Gamma(t)\Delta t \end{aligned}$$

$$\begin{aligned} \Gamma(t) = {} & P(t|t)[f_x(\hat{x}(t|t),t)-S(t)h_x(\hat{x}(t|t),t)]^T+[f_x(\hat{x}(t|t),t)- \\ & -S(t)h_x(\hat{x}(t|t),t)]P(t|t)+\bar{Q}(t) \end{aligned}$$

$$\begin{aligned} a(x_{k+1},z_{k+1}) &\equiv a(x(t+\Delta t),y(t+\Delta t)) \\ &= h_x^T(x(t+\Delta t),t+\Delta)R^{-1}(t+\Delta t)[\dot{y}(t+\Delta t)-B(t+\Delta t)y(t+\Delta t)- \\ &\quad -h(x(t+\Delta t),t+\Delta t)]\Delta t \end{aligned}$$

$$\begin{aligned} K_{k+1|k} &\equiv K(t+\Delta t|t) \\ &= \{P(t|t)+[P(t|t)a_x(\hat{x}(t+\Delta t|t),y(t+\Delta t))P(t|t)+\Gamma(t)]\Delta t\}h_x^T(\hat{x}(t+\Delta t|t), \\ &\quad t+\Delta t)R^{-1}(t+\Delta t) \end{aligned} \tag{38}$$

Utilizing eq. (38), we obtain from Eqs. (19) and (21), as $\Delta t \to 0$,

$$\begin{aligned} \dot{\hat{x}}(t|t) = {} & f(\hat{x}(t|t),t)+[P(t|t)h_x(\hat{x}(t|t),t)+W(t)]R^{-1}(t)[\dot{y}(t)-B(t)y(t)- \\ & -h(\hat{x}(t|t),t)] \end{aligned} \tag{39}$$

$$\hat{x}_o = \bar{x}_o+X_o h_x^T(\bar{x}_o,t_o)[h_x(\bar{x}_o,t_o)X_o h_x^T(\bar{x}_o,t_o)+R_o]^{-1}[z_o-h(\bar{x}_o,t_o)]$$

$$\dot{P}(t|t) = P(t|t)f_x^T(\hat{x}(t|t),t)+f_x(\hat{x}(t|t),t)P(t|t)+G(t)Q(t)G^T(t)-$$
$$-[P(t|t)h_x^T(\hat{x}(t|t),t)+W(t)]R^{-1}(t)[P(t|t)h_x^T(\hat{x}(t|t),t)+W(t)]^T \quad (40)$$
$$P_o = X_o-X_o h_x^T(\bar{x}_o,t_o)[h_x(\bar{x}_o,t_o)X_o h_x^T(\bar{x}_o,t_o)+R_o]^{-1}h_x(\bar{x}_o,t_o)X_o,$$

where

$$W(t) = G(t)Q(t)G^T(t)H^T(t)+G(t)C(t)F^T(t). \quad (41)$$

Equations (39) and (40) are the required continuous-time filter equations. These equations agree with the results obtained in [9].

(b) Continuous-time Smoother Equations

We have the following analogues to within the first-order terms in Δt:

$$b(x_k,\hat{x}_{k+1|N},z_k) \equiv b(x(t),\hat{x}(t+\Delta t|t_1),y(t))$$
$$= [I+f_x(x\,t),t)\Delta t-S(t)-s(t)h_x(x(t),t)\Delta t]^T(\frac{\bar{Q}(t)}{\Delta t})^{-1} \cdot$$
$$\cdot\{\hat{x}(t+\Delta t|t_1)-x(t)-f(x(t),t)\Delta t-S(t)[\dot{y}(t)-B(t)y(t)-$$
$$-h(x(t),t)]\Delta t\}$$

$$b_x(\hat{x}(t|t),\hat{x}(t+\Delta t|t_1),y(t)) = -(\frac{1}{\Delta t})M^T(t)\bar{Q}^{-1}(t)[I+f_x(\hat{x}(t|t),t)\Delta t-$$
$$-S(t)h_x(\hat{x}(t|t),t)\Delta t]+\beta_x(\hat{x}(t|t),\hat{x}(t+\Delta t|t_1),y(t))$$

$$M(t) = [I+f_x(\hat{x}(t|t),t)\Delta t-S(t)-S(t)h_x(\hat{x}(t|t),t)\Delta t]$$

$$\beta_x(\hat{x}(t|t),\hat{x}(t+\Delta t|t_1),y(t)) = \frac{\partial}{\partial x}[f_x(x(t),t)-S(t)h_x(x(t),t)]\bar{Q}^{-1}(t)[\hat{x}(t+\Delta t|t_1)-$$
$$-\hat{x}(t+\Delta t|t)]\}\Big|_{x=\hat{x}(t|t)}$$

$$A_k^{-1} \equiv A^{-1}(t) = I+\Psi(t)\Delta t$$

$$\Psi(t) = f_x(\hat{x}(t|t),t)-S(t)h_x(\hat{x}(t|t),t)+\bar{Q}(t)[M^T(t)]^{-1}[P^{-1}(t|t)-$$
$$-\beta_x(\hat{x}(t|t),\hat{x}(t+\Delta t|t_1),y(t))]$$

$$A^{-1}(t)P(t|t)[A^{-1}(t)]^T = P(t|t)+\Psi(t)P(t|t)\Delta t+P(t|t)\Psi^T(t)\Delta t$$
$$A^{-1}(t)P(t|t_1)[A^{-1}(t)]^T = P(t|t_1)+\Psi(t)P(t|t_1)\Delta t+P(t|t_1)\Psi^T(t)\Delta t \quad (42)$$

Utilizing Eq. (42), we obtain from Eqs. (27) and (30) for all $t\varepsilon[t_o,t_1]$, as $\Delta t\to 0$,

$$\dot{\hat{x}}(t|t_1) = f(\hat{x}(t|t),t)+D(t)[\hat{x}(t|t_1)-\hat{x}(t|t)]+S(t)[\dot{y}(t)-B(t)y(t)-h(\hat{x}(t|t),t)] \quad (43)$$

$$\dot{P}(t|t_1) = P(t|t_1)D^T(t)+D(t)P(t|t_1)+[P(t|t)\beta_x^T(\hat{x}(t|t),\hat{x}(t|t_1),y(t))+$$
$$+\beta_x(\hat{x}(t|t),\hat{x}(t|t_1),y(t))P(t|t)-I]\bar{Q}(t), \quad (44)$$

where

$$D(t) = f_x(\hat{x}(t|t),t)-S(t)h_x(\hat{x}(t|t),t)+\bar{Q}(t)[P^{-1}(t|t)-\beta_x(\hat{x}(t|t),\hat{x}(t|t_1),y(t))]. \quad (45)$$

The initial conditions for Eqs. (43) and (44) are, respectively, $\hat{x}(t_1|t_1)$ and $P(t_1|t_1)$, which are the filter outputs obtained from Eqs. (39) and (40) at the terminal time t_1.

We note that it is not necessary to store the matrix $P^{-1}(t|t)$ in Eq. (45) and hence in Eq. (44). By defining $\bar{P}(t) \triangleq P^{-1}(t|t)$, we immediately obtain from Eq. (40),

$$\begin{aligned}\dot{\bar{P}}(t) &= -f_x^T(\hat{x}(t|t),t)\bar{P}(t)-\bar{P}(t)f_x(\hat{x}(t|t),t)-\bar{P}(t)G(t)Q(t)G^T(t)\bar{P}(t)+\\ &+[h_x^T(x(t|t),t)+\bar{P}(t)W(t)]R^{-1}(t)[h_x^T(\hat{x}(t|t),t)+\bar{P}(t)W(t)]^T\\ \bar{P}(t_1) &= P^{-1}(t_1|t_1)\end{aligned} \tag{46}$$

for all $t\in[t_o,t_1]$. Eq. (46) can be computed to provide the required matrix $\bar{P}(t)$ in Eq. (44). The only matrix inversions needed are $P^{-1}(t_1|t_1)$ at the terminal time t_1 and $\bar{Q}^{-1}(t)$; the latter requires that $\bar{Q}(t)$ be positive-definite. We note that when the system is linear, $\beta_x(\hat{x}(t|t),\hat{x}(t|t_1),y(t))=0$ for all $t\in[t_o,t_1]$, and no inverse of $\bar{Q}(t)$ is needed.

APPENDIX

The following table is required to establish the analogues between the discrete-time and the continuous-time models (see also [8] and [12]).

Discrete-time Model	Continuous-time Model
k, N $(N > k)$	t, t_1 $(\Delta t > 0,\ t_1 > t)$
x_k, x_{k+1}	$x(t)$, $x(t+\Delta t)$
$f(x_k,k)$	$x(t)+f(x(t),t)\Delta t$
G_k, F_k	$G(t)\Delta t$, $F(t)\Delta t$
H_k	$H(t)$
Q_k, N_k, C_k	$Q(t)/\Delta t$, $N(t)/\Delta t$, $C(t)/\Delta t$
B_k	$I+B(t)\Delta t$
z_k	$[\dot{y}(t)-B(t)y(t)]\Delta t$
$h(x_k,k)$	$h(x(t),t)\Delta t$
	$=H(t+\Delta t)f(x(t),t)\Delta t+$
	$+[\dot{H}(t)-B(t)H(t+\Delta t)]x(t)\Delta t$
R_k, W_k, S_k	$R(t)\Delta t$, $W(t)\Delta t$, $S(t)$

REFERENCES

[1] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. J. Basic Eng., Ser. D, 83 (1961), 95-108.

[2] A. E. Bryson and D. E. Johansen. Linear filtering for time-varying systems using measurements containing colored noise. IEEE Trans. Automatic Control, AC-10 (1965), 4-10.

[3] R. S. Bucy. Optimal filtering for correlated noise. J. Math. Anal. Appl., 20 (1967), 1-8.

[4] R. S. Bucy and P. D. Joseph. "Filtering for Stochastic Processes with Application to Guidance." Interscience, New Gork, 1968.

[5] E. B. Stear and A. R. Stubberud. Optimal filtering for Gauss-Markov noise. Int. J. Control, 8 (1968), 123-130.

[6] A. E. Bryson and L. J. Henrikson. Estimation using sample data containing sequentially correlated noise. J. Spacecraft, 5 (1968), 662-665.

[7] A. E. Bryson and Y. C. Ho. "Applied Optimal Control." Blaisdell, Waltham, Massachusetts, 1969.

[8] R. E. Kalman. New methods and results in linear prediction and filtering theory. In "Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability" editoed by J. L. Bogdanoff and F. Kozin, Wiley, New York, 1963, 270-388.

[9] N. K. Loh. Nonlinear filtering for Gauss-Markov observation noise. 1969 (to appear).

[10] H. E. Rauch, F. Tung and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. AIAA J., 3 (1965), 1445-1450.

[11] J. S. Meditch. Sequential estimation for discrete-time nonlinear systems. Boeing document D1-82-0830, Boeing Scientific Research Laboratories, Seattle, Washington, (1969).

[12] J. S. Meditch. Formal algorithms for continuous-time nonlinear filtering and smoothing. Boeing document D1-82-0838, Boeing Scientific Research Laboratories, Seattle, Washington, (1969).

[13] T. W. Anderson. "An Introduction to Multivariate Statistical Analysis." Wiley, New York, 1958.

[14] K. Ito. On stochastic differential equations. Memoirs of the American Mathematical Society, 4, (1951).

[15] J. L. Doob. "Stochastic Processes." Wiley, New York, 1953.

[16] K. Ito. On a formula concerning stochastic differentials. J. Nagoya Math., 3, (1951), 55-65.

ON REALIZATION OF A CIRCUIT MATRIX

SHU-GAR CHAN
Naval Postgraduate School
Monterey, California

JAMES F. NALE
Naval Ships Engr. Center
Washington, D.C.

ABSTRACT

This paper presents a method to test any given matrix for realizability as a circuit matrix of a nonoriented nonseparable graph. The method also provides a realization if the matrix under test is realizable.

INTRODUCTION

The problem of realization of a given matrix as a circuit matrix has been studied by a number of investigators. Methods given include the combinatorial approach and the growing of a graph using constricted matrices; they range from the synthesis of a given realizable fundamental circuit matrix to the realization of special types of matrices. Some representations of these methods are given in referenses [1] through [5].

REALIZATION PROCEDURE

Ash and Kim [4] worked out a systematic procedure to synthesize a realizable circuit matrix for any matrix of which the corresponding tree branches form a maximal path of a graph. If the columns of a circuit matrix B corresponding to the tree branches are arranged so that no two 1's are separated by one or more 0's, then the branches of the tree taken in the order as they appear in B form a maximal path of the graph, which is then realized.

The method of Ash and Kim is now extended to the general case of nonoriented nonseparable graph. We shall first outline the realization procedure below.

Step 1: Determination of the Number of Vertices.

By Jordan's elimination method a given matrix B may be reduced so that the number of vertices v of the corresponding graph G, if it exists, can be determined.

Step 2: Obtain the Reduced Circuit Matrix Br.

Generate all possible circuits by modulo-2 addition of all combinations of the independent rows. Eliminate all edge-disjoint unions of circuits. Use series combination and parallel combination to combine respectively all series edges and parallel edges. This operation results in Br.

Step 3: The First Subgraph.

(a) The Hamilton Circuit

Select a row of Br with the largest number of non-zero entries. This is a Hamilton circuit of some subgraph G_1.

(b) Tree of G_1

Use all except one of the edges of this Hamilton circuit as the tree T_1 of this subgraph G_1.

(c) Completing the Submatrix B_1

Select all rows of B_r that have exactly one edge not contained in T_1. Form a submatrix B_1 using these rows.

(d) Realization of the First Subgraph

Arrange the columns of B_1 corresponding to the tree so that no two 1's are separated by one or more 0's in any row of this submatrix. If this is not possible, the subgraph is not be realizable.

(e) Simplify B_r

Delete all rows of B_r that contain a chord of the tree in the subgraph that has been realized.

Step 4: The Second Subgraph.

(a) The Hamilton Circuit

Select a row of B_r with the largest number of nonzero entries that has at least two edges in common with the tree of the subgraph that has been realized. This is a Hamilton circuit of some subgraph G_2.

(b) Tree of G_2

Use all except one of the edges of this Hamilton circuit as the tree T_2 of G_2.

(c) Completing the Submatrix.

Select all rows of B_r that have exactly one edge not contained in the tree. Form a submatrix B_2 using these rows.

(d) Realization of the Second Subgraph.

Arrange the columns of B_2 corresponding to the tree of this submatrix so that no two 1's are separated by one or more 0's in any row of this submatrix.

(e) Joining the Subgraphs

Arrange the columns of the common edges of B_1 and B_2 so that they appear in the same order.

(f) Simplify B_r

Delete all rows of B_r that contain a chord of the tree in the subgraph that has been realized.

(g) Draw the Partial Tree

When the subgraphs are joined, there are three possible configurations of the tree, or partial tree, thus far obtained (see Fig. 4).

For case 1, all columns contained in both trees can be arranged so that no two 1's are separated by one or more 0's in any row when the columns of the second submatrix are placed adjacent to the columns of the first. For cases 2 and 3, the tree configuration must be carried forward.

Step 5: Subsequent Subgraphs.

(a) The Hamilton Circuit

Select a row of B_r with the largest number of non-zero entries that has at least two edges in common with a path in the partial tree obtained above. This is a Hamilton circuit of some subgraph.

(b) Tree of the Subsequent Subgraph

Use all except one of the edges of the Hamilton circuit as the tree of this subgraph.

(c) Completing Submatrix

Select all rows of B_r that have exactly one edge not contained in the tree. Form a submatrix with these rows.

(d) Realization of Subsequent Subgraphs

Arrange the columns of submatrix in (c) corresponding to the tree in (b) so that no two 1's are separated by one or more 0's in any row of this submatrix.

(e) Joining the Subgraphs

Arrange the columns of the common edges of this and previously realized submatrices so that they appear in the same order.

(f) Simplify B_r

Delete all rows of B_r that contain a chord of the tree in the subgraph that has been realized.

(g) Draw the Partial Tree

Step 6: Realization of the Graph

(a) Realization of B_r

Repeat the procedure outlined in Step 5 until there are no rows left in B_r. At this point, there will be v-1 edges in the tree and the realization is complete.

(b) Reinsertion of Series and Parallel edges.

The graph is completed by reinserting the series and parallel edges that were removed in Step 2.

REALIZABILITY

The procedure given in the last section will either lead to a realization or find that the given circuit matrix is not realizable. From references [4] and [5], we realize that if (v-1) edges of the Hamilton circuit are chosen as the branches of the tree, then the condition for the realizability of the given circuit matrix (containing a Hamilton circuit) is that the columns of the tree can be arranged so that no two 1's are separated by one or more 0's in any independent row of the matrix.

Extending the method of realizing submatrices containing Hamilton circuits the procedure in the last section will lead to a realization of any circuit matrix when each step is completed. Otherwise, if part (d) of step 2 or 3 or 4 cannot be completed, the given matrix is not realizable. A nonrealizable "circuit matrix" will contain one or more of the following submatrices:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

EXAMPLES

Example 1

Let a given circuit matrix be reduced to the form in (2).

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	1	1	1	0	1
0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	1
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0

(2)

$$\begin{bmatrix} 0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&1&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&1&0 \end{bmatrix}$$

The third and fourth rows imply that edges c and d are parallel edges, replace them with edge c'. The seventh and eighth rows imply that edges g and h are parallel edges, replace them with edge g'. The tenth and eleventh rows imply that edges i and k are parallel edges, replace them with edge j'.

Since edge a or o does not appear in a row without the other also appearing, edges a and o are series edges, replace them with edge a'. The same argument holds for edges e and s as well as for edges t and x. Replace edges e and s with edge d'. Replace edges t and x with edge t .

This matrix can be put into the following form:

$$\begin{array}{c} \begin{matrix} a' & b & c' & e' & f & g' & i & j' & l & m & n & p & q & r & t' & u & v & w \end{matrix} \\ \begin{bmatrix} 1&0&0&0&0&0&0&0&1&1&0&0&0&0&1&1&1&0 \\ 0&1&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0&0&1&1&0&1&1&1&0 \\ 0&0&0&0&1&0&0&0&0&0&0&1&1&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&1 \end{bmatrix} \end{array}$$

There are no series edges or parallel edges in the graph corresponding to this circuit matrix. Modulo-2 addition of all combinations of the rows of this circuit matrix, after elimination of all edge-disjoing unions of circuits, will yield B_r. B_r is shown in figure 1.

Select row 19 as a Hamilton circuit for the first subgraph. The tree for this subgraph is chosen to be edges c', e', f, g', l, n, and q. The rows of B_r that have exactly one edge not in this tree are 19, 33, 51, 52, and 54. This submatrix can be put into the following form:

$$\begin{array}{c} \begin{matrix} & a' & b & m & p & r & l & c' & n & q & g' & f & e' \end{matrix} \\ \begin{matrix} 19 \\ 33 \\ 51 \\ 52 \\ 54 \end{matrix} \begin{bmatrix} 1&0&0&0&0&1&1&1&1&1&1&1 \\ 0&1&0&0&0&1&1&0&0&0&0&0 \\ 0&0&1&0&0&0&1&1&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&1&1&0 \\ 0&0&0&0&1&0&0&0&1&1&0&0 \end{bmatrix} \end{array}$$

It is seen that this submatrix meets the cuiterion for realizability and the tree is a path of the subgraph consisting of the edges 1, c', n, q, g', f, and e' in that order. After the rows of B_r containing edges a', b, m, p, and r have been deleted, the following rows remain in B_r: 43, 44, 45, 46, and 57.

Select row 45 as a Hamilton circuit for the second subgraph. The tree for this subgraph is chosen to be edges e', f, g', j', q, t', and v. The rows of B_r that have exactly one edge not contained in this tree are 45, 57, and 44. This submatrix can be put into the following form:

	i	w	u	q	g'	f	e'	t'	v	j'
45	1	0	0	1	1	1	1	1	1	1
57	0	1	0	0	0	0	0	0	1	1
44	0	0	1	1	1	1	1	1	1	0

It is seen that this submatrix meets the criterion for realizability and the tree is a path of the subgraph consisting of the edges q, g', f, e', t', v, and j' in that order. It is noted that the edges common to both submatrices appear in the same order. Since the common edges appear on the ends of the paths in both subgraphs, the tree obtained when the two subgraphs are joined will be a path in the graph. The matrix for the combination of the two submatrices can be put into the following form:

	a'	b	m	p	r	i	w	u	1	c'	n	q	g'	f	e'	t'	v	j'
19	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0
33	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
51	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
52	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
54	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
45	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1
57	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
44	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0

Since there are no rows in B_r when rows containing edges i, w, and u are deleted, this matrix is a fundamental circuit matrix of the graph corresponding to B_r. The tree of the graph is a path. The realized graph is shown in Fig. 2.

Figure 3 shows the completed graph after the series and parallel edges that were removed by combination are reinserted.

Example 2

Let the given circuit matrix be

$$
\begin{array}{c}
\begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \end{array} \\
\begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
\qquad (3)
$$

Using Jordan's elimination method, (3) can be reduced to the following form:

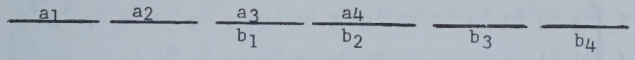
$$
\begin{array}{c}
\begin{array}{ccccccccc} a & b & c & d & f & e & g & h & i \end{array} \\
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1
\end{bmatrix}
\end{array}
\qquad (4)
$$

The first row implies that edges a and g are parallel edges, replace them with edge a'. The second and fourth rows imply that edges b and d are parallel edges, replace them with edge b'. The resulting matrix can be put into the following from:

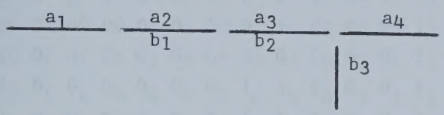
$$
\begin{array}{c}
\begin{array}{ccccccc} a' & b' & c & f & e & h & i \end{array} \\
\begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 0 & 1
\end{bmatrix}
\end{array}
\qquad (5)
$$

The matrix in (5) contains the submatrix

(Case 1)



(case 2)



(Case 3)

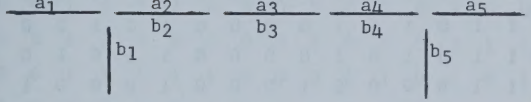


Figure 4

$$\begin{array}{c} f \quad e \quad h \\ \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \end{array}$$

which is not a realizable submatrix so that the given circuit matrix is not realizable.

REFERENCES

[1] Seshu, S., "Topological Considerations in the Design of Driving Point Functions", IRE Trans. on Circuit Theory, pp. 193-214, Oct. 1955.

[2] Parker, S.R. and Lohse, H.J., "A Direct Procedure for the synthesis of Network Graphs from a Given Fundamental Loop or Cut Set Matrix", IEEE Trans. on Circuit Theory, May 1969.

[3] Fu, Y., "Realization of Circuit Matrices", IEEE Trans. on Circuit Theory, Vol. CT-12, No. 4, pp. 604-607, Dec. 1965.

[4] Ash, R. B. and Kim, W. H., "On Realizability of a Circuit Matrix", IRE Trans. on Circuit Theory, Vol. CT-6 pp. 219-223, June 1959.

[5] Gould, R., "Graphs and Vector Spaces", Journal of Math. and Physics, Vol. 37, 1958, pp. 193-214.

	a'	b	c'	e'	f	g'	i	j'	l	m	n	p	q	r	t'	u	v	w
1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
3	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
4	1	1	1	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0
5	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
6	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
7	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0
8	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1
9	1	1	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0
10	1	1	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0
11	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0
12	1	1	0	1	0	1	0	0	0	0	1	1	0	1	0	0	0	0
13	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1
14	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0
15	1	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0
16	1	1	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1
17	1	0	1	1	0	0	0	0	1	0	1	1	1	0	0	0	0	0
18	1	0	1	1	1	0	0	0	1	0	1	0	0	1	0	0	0	0
19	1	0	1	1	1	1	0	0	1	0	1	0	1	0	0	0	0	0
20	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	0	0
21	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	1
22	1	0	1	0	0	0	0	0	1	0	1	0	0	0	1	1	1	0
23	1	0	1	0	0	0	1	1	1	0	1	0	0	0	1	0	1	0
24	1	0	1	0	0	0	0	1	1	0	1	0	0	0	1	1	0	1
25	1	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0
26	1	0	0	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0
27	1	0	0	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0
28	1	0	0	1	0	1	0	0	1	1	0	1	0	1	0	0	0	0
29	1	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	1
30	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0
31	1	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	1	0
32	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	0	1
33	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
34	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0

Figure 1

	a'	b	c'	e'	f	g'	i	j'	l	m	n	p	q	r	t'	u	v	w
35	0	0	0	1	0	0	1	0	0	0	0	1	1	0	1	0	0	1
36	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	1	0
37	0	0	0	1	0	0	1	1	0	0	0	1	1	0	1	0	1	0
38	0	0	0	1	0	0	0	1	0	0	0	1	1	0	1	1	0	1
39	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	0	0	1
40	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0
41	0	0	0	1	1	0	1	1	0	0	0	0	0	1	1	0	1	0
42	0	0	0	1	1	0	0	1	0	0	0	0	0	1	1	1	0	1
43	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	0	0	1
44	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	1	1	0
45	0	0	0	1	1	1	1	1	0	0	0	0	1	0	1	0	1	0
46	0	0	0	1	1	1	0	1	0	0	0	0	1	0	1	1	0	1
47	0	0	0	1	0	1	1	0	0	0	0	1	0	1	1	0	0	1
48	0	0	0	1	0	1	0	0	0	0	0	1	0	1	1	1	1	0
49	0	0	0	1	0	1	1	1	0	0	0	1	0	1	1	0	1	0
50	0	0	0	1	0	1	0	1	0	0	0	1	0	1	1	1	0	1
51	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
52	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
53	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0
54	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
55	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0
56	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1
57	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1

Figure 1--continued

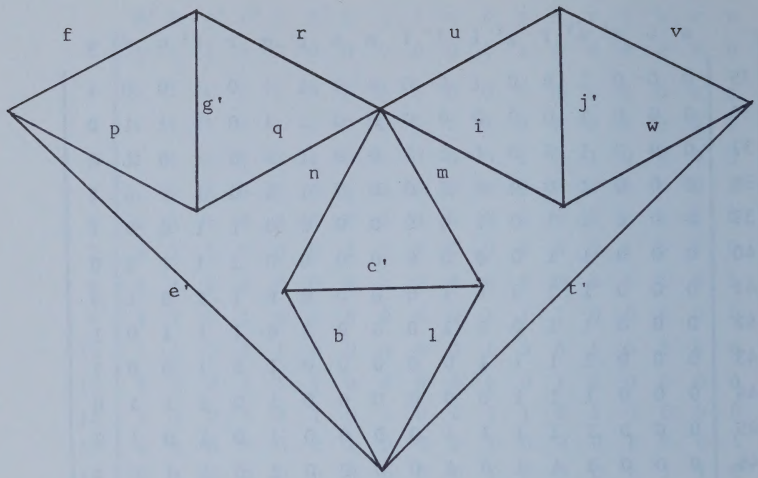


Figure 2

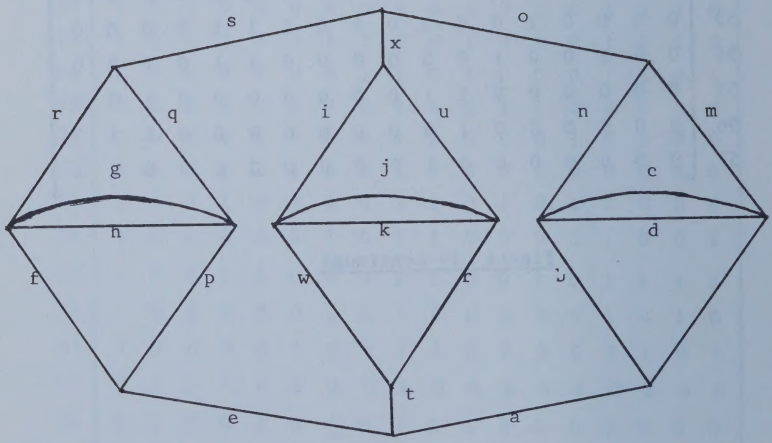


Figure 3

REARRANGEABILITY OF TRIANGULAR CONNECTION NETWORKS

M. C. Paull
Rutgers University
New Brunswick, New Jersey

Abstract

In this paper the relation between the problem of setting up connections in a connection network which is designed to connect any set of pairs of inputs, and a problem in graph theory is established. The graph theory problem is then solved. The main result in graph theory terms is a constructive proof that no more than $\left\lfloor \frac{3n}{2} \right\rfloor$ colors are necessary to color the branches of a graph each of whose nodes is of degree n or less. In connection networks terms, the result is an algorithm for rearranging connections in the network so as to make the network effectively non-blocking.

Index Terms: Connection Networks, Non-blocking, Graph Theory, Rearrangeability, Permutation Networks, Telephone Network, Graph coloring.

Introduction

In a telephone office a local call is completed by connecting the telephone lines of the two parties involved through a set of relay contacts in an array of such contacts, called the connection network.

A simplified example of such a network, designed to allow the connection of any set of disjoint pairs of four input lines is illustrated below:

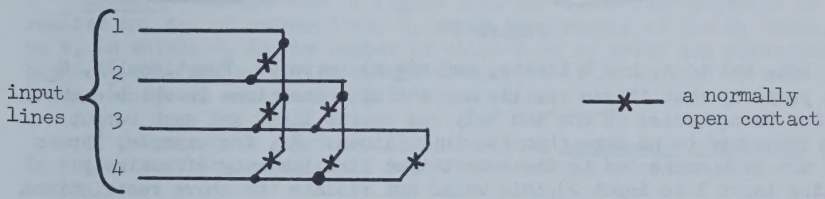


Figure 0

This network is a single-stage triangular network. Such networks with n inputs require one contact for every pair of input lines, or $n(n-1)/2$ contacts. For any reasonable number of inputs this is very costly. On the other hand this network is strongly non-blocking, that is, any time it is necessary to connect any two idle input lines together there will be a contact available. Furthermore, this type of network is relatively easy to control.

There are a variety of types of connection networks available which offer the designer various combinations of the three key factors: blocking characteristic, contact count per input, and control complexity.

In this paper we will consider a class of multi-staged connection networks which are rearrangeably non-blocking, or more simply rearrangeable. A rearrangeable network is one in which any two presently unconnected lines can be connected, perhaps requiring the rearrangement or reconnection through

other paths of previously established connections, but without finally destroying any of these connections. This can be done no matter which other pairs of lines are already connected. Note that a strongly non-blocking network is a rearrangeable network in which no rearrangement or reconnections will ever be necessary. For a network which requires rearrangement contact cost is generally less although control is more complex than for a strongly non-blocking network.

Our purpose is to show that a particular class of connection networks is rearrangeable. An algorithm for carrying out these arrangements is implicitly included in the proof that the network in question is rearrangeable.

The Network and an Associated Graph Theory Prober

A multi-stage triangular connection network is pictured below:

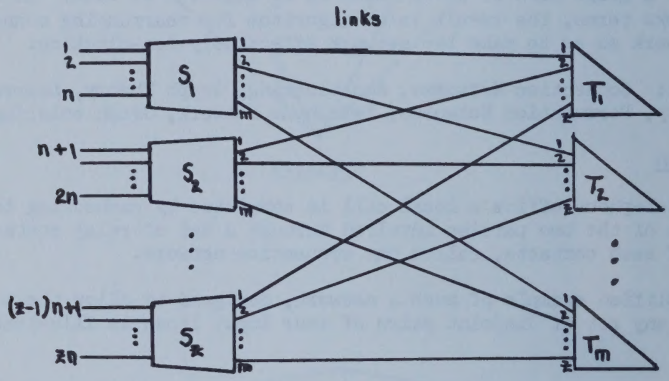


Figure 1

Each S_i box, i=1 to z, has n inputs, and m($\geq$ n) outputs. Functionally, S_i has the property that it can realize any set of connections in which each input line is connected to one and only one output line, and each output line is connected to no more than two input lines. If, for example, inputs 1 and 2 are both connected to the same output line (thereby effectively connecting input 1 to input 2), this would not violate the above restrictions.

Each T_i has the property that it can realize any set of connections in which each input is connected to at most one other input (such a set of connections is called legitimate). T_i is then, by definition, a rearrangeable network. Since all its inputs are on one-side, we will call it one-sided rearrangeable when we wish to call attention to this feature. The network of Figure 0 is an example of a one-sided rearrangeable network.

The composite network of Figure 1, with its nz inputs, has been shown to be one-sided rearrangeable by Clos [1] if m=2n - 1. In fact, he showed that if m = 2n - 1, then the network of Figure 1 is strongly non-blocking.

Here we will show:

Theorem 1: The network of figure 1 is one-sided rearrangeable if m$\geq\lfloor\frac{3n}{2}\rfloor$,* and m must be at least $\lfloor\frac{3n}{2}\rfloor$ or the network is not one-sided rearrangeable.

* $\lfloor x\rfloor$ is x if x is an integer and the next lowest integer if x is not an integer.

In general, with $\left\lfloor \frac{3n}{2} \right\rfloor \leq m < 2n-1$, the addition of a new connection when the network is in a legitimate state will require rearranging the switches involved in connections already set up. Our proof that with $m \geq \left\lfloor \frac{3n}{2} \right\rfloor$ the network of figure 1 is one-sided rearrangeable consists of showing how, when a new connection is to be added, the existing connections can be rearranged to make this possible.

As a first step toward that proof, we will examine what the definition of one-sided rearrangeable implies about the network of figure 1.

To connect any input, say 1 of S_i, to any input, say 2 of S_j; i,j=1 to z (i can = j), it is only necessary that there be unoccupied links, one an output of S_i, say link a; and one an output of S_j, say link b; with both connected to the same second stage switch T_a (note if $S_i = S_j$ then link a must = link b). This is adequate because, by the nature of the boxes labelled S_i, i=1 to z, a connection can always be made from any unoccupied input (1 for example) through S_i to any unoccupied output of S_i (link a for example). Similarly a connection can be made from 2 through S_j to link b. Finally, from the nature of the T_i boxes, i=1 to m, a connection can always be made between any two unoccupied inputs as links a and b are in this case. In other words, we only need be able to make a connection <u>from</u> box S_i to box S_j to be able to make a connection from any input on box S_i to any input on box S_j. In the special case in which we wish to connect an input on S_i to another input on S_i, the above implies the use of a single unoccupied output on S_i. Therefore, it will only be necessary to consider possible connections between the <u>outputs</u> of S_i and S_j, i,j=1 to z, without being concerned with precisely which inputs of S_i and S_j are involved. Recalling the definition of one-sided rearrangeable, it follows easily that:

<u>Lemma 1</u>: The network of Figure 1 is one-sided rearrangeable if it can realize any set of connections, C, among the outputs of the S_i boxes, i=1 to z, in which; o_i is the number of outputs of S_i which are connected to boxes $S_j, j \neq i$, and s_i is the number of outputs of S_i to each of which two inputs of S_i are connected and in which <u>$n \geq 2s_i + o_i$</u>.

If C meets this condition it is called a <u>legitimate connection</u> set. In the network the connections specified in C are established in the T_i boxes.

Now we will convert the problem of how these legitimate connections sets are to be realized to a problem of coloring the branches of a graph.

We can represent each legitimate set of connections, C, of Lemma 1 by a graph G_C with z nodes labelled $S_1, S_2 \ldots S_z$. Each node corresponds to an S_i box, i=1 to z, of Figure 1. For each connection in C <u>from</u> box S_i to S_j, i,j=1 to z, there is a branch in the graph G_C from node S_i to S_j. For example, the graph of Figure 2 corresponds to a set of connections in which z=3, n=4. There is a connection from box S_1 to itself, three from S_2 to S_3, one from S_1 to S_2 and one from S_1 to S_3.

Such a connection graph for a legitimate set of connections in the triangular network of Figure 1 must have:

1. z nodes
2. if o_i is the number of branches from a node S_i to nodes other than S_i and s_i is the number of branches from S_i to itself then $n \geq o_i + 2s_i$.

Any graph with the above restrictions represents a legitimate set of connections for the network of Figure 1. The set of all such graphs is called S(z,n).

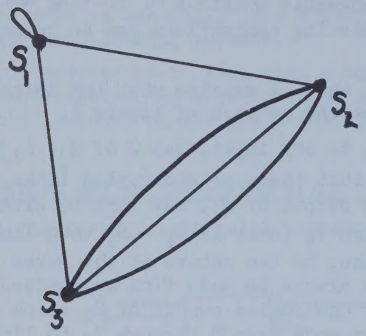


Figure 2

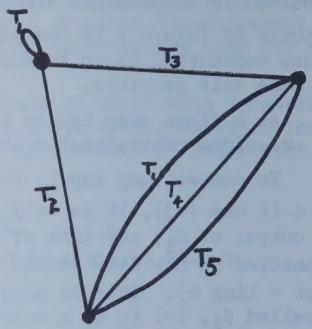


Figure 3

Now if we can assign one of a set of colors, say $T_1T_2...T_m$, to each branch of a graph in such a way that no node has incident upon it two distinct branches (a branch from a node to itself is 1 branch even though both ends are incident on the node) with the same color, we say that the graph can be m-colored. Thus for example, in Figure 3, we show that the graph of Figure 2 can be colored with 5 colors. It is easy to see that if all the graphs in S(z,n) can be m-colored, then the network of Figure 1 is one-sided complete. Each graph in S(z,n) corresponds to a legitimate connection set; m-coloring the graph corresponds to a way of realizing that connection set. If the color T_k appears on a branch from a node S_i to node S_j, i,j=1 to z, $i \neq j$, then a connection is to be made in the network in box T_k from a link from S_i to a link from S_j. This can always be done because we are guaranteed that there will be no more than 1 branch from S_i to S_j colored T_k, and in the network, there is one link from each of the boxes S_i and S_j to box T_k, so the required connection can be made in T_k with no interference with any other connection. If there is a branch from node S_i to itself colored T_a, this implies that two inputs to S_i are to be connected together by both being connected to the output of S_i which connects to box T_a. This can always be accomplished since this output of S_i will not be required for any other connection; the color T_a can only be the color of one branch incident on node S_i. Thus each graph in S(z,n) which can be m-colored can be realized in the network of Figure 1. Also, it is easy to show that if a legitimate connection set can be realized, its corresponding graph can be m-colored.

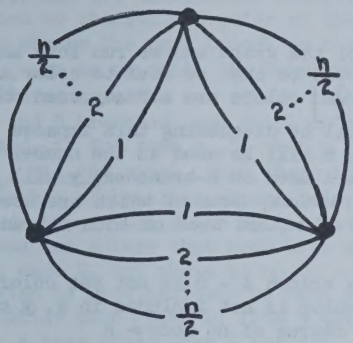
Lemma 2: All the graphs in S(z,n) can be m-colored if the network of Figure 1 is one-sided complete or rearrangeable.

Theorem 1, our main result, follows immediately from Lemma 1, Lemma 2, and the following two key theorems.

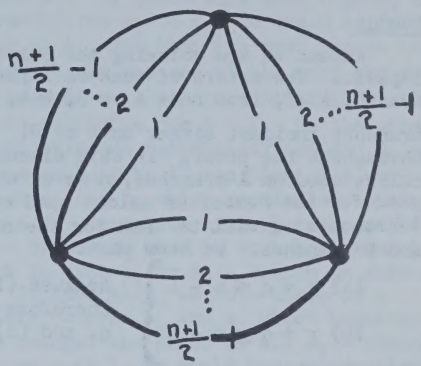
Theorem 2a: If $z \geq 3$, then there are graphs in S(z,n), that can be m-colored only if $m \geq \left\lceil \frac{3n}{2} \right\rceil$.

Proof: We can prove this by exhibiting graphs having 3 nodes which require $\left\lfloor \frac{3n}{2} \right\rfloor$ colors having:

For n even



For n odd



# of colors = # of branches = $3\,\frac{n}{2}$, n even; = $3\left(\frac{n+1}{2}\right) - 2$, n odd

QED

Figure 4

Theorem 2b: For any z,n, each graph in S(z,n), can be m-colored with m no greater than $\left\lfloor \frac{3n}{2} \right\rfloor$.

The rest of this paper is devoted to a constructive proof of this theorem. The proof will show how to color the branches of any given graph in S(z,n) with no more than $\left\lfloor \frac{3n}{2} \right\rfloor$ colors, one at a time - and sometimes recoloring previously colored branches so as to make possible the legitimate coloring of an as-yet-uncolored branch. This construction when translated to network terms, provides algorithms which can be used to add a new connection between inputs of the network of Figure 1, rearranging existing connections where necessary.

Proof of Main Theorem

First we can eliminate from consideration all graphs in S(z,n) with self-loops, i.e. branches which go from a node to the same node. Because if one removes all the self-loops and can then color the remaining graph with no more than $\left\lfloor \frac{3n}{2} \right\rfloor$ colors, then after putting the self-loops back, each on its unique node, one can color each self-loop using one of the $\left\lfloor \frac{3n}{2} \right\rfloor$ colors not already used for branches incident on that node. Since no more than n branches are incident on any node, there are always colors available for this purpose.

So we will consider only those graphs in S(z,n) which have no self-loops. The following additional notation is used in the proof.

We use small letters for colors, capital letters for nodes.

An X-branch is a branch incident on the node X.

An x-branch is a branch whose color is x. where { X and Y are any capital letter; x is any small letter }

X - Y is a branch from node X to node Y.

The remainder of the proof is organized in a number of titled sections, the first of which is:

Bounds

Assume we are coloring the branches of the graph and we run into an impasse. The nature of such an impasse would be that we wish to color a branch, A -B, from node A to B, but, all $\left\lfloor \frac{3}{2}n \right\rfloor$ colors are already used on branches incident either on A or B. We will be discussing this impasse throughout the proof. In this discussion, x will be used as the number of colors used on A-branches, none of which are used on B-branches; y will be used for the number of colors used on B-branches, none of which are used on A-branches; q will be used for the number of colors used on both A-branches and B-branches. We know that:

1a) $x + q \leq n - 1$

1b) $y + q \leq n - 1$

} because (1) the branch A - B is not yet colored therefore its color is not included in x, y or q, and (2) the degree of no node $> n$

also clearly:

2) $x + y + q = \left\lfloor \frac{3}{2}n \right\rfloor$

and therefore subtracting 1a, and 1b from 2 we get respectively:

3a) $x \geq x_{min} = \begin{cases} \frac{n}{2} + 1 & n \text{ even} \\ \frac{n-1}{2} + 1 & n \text{ odd} \end{cases}$

3b) $y \geq y_{min} = \begin{cases} \frac{n}{2} + 1 & n \text{ even} \\ \frac{n-1}{2} + 1 & n \text{ odd} \end{cases}$

The initial impasse

These bounds will be used later in the discussion.

We can picture the impasse situation we are discussing as follows:

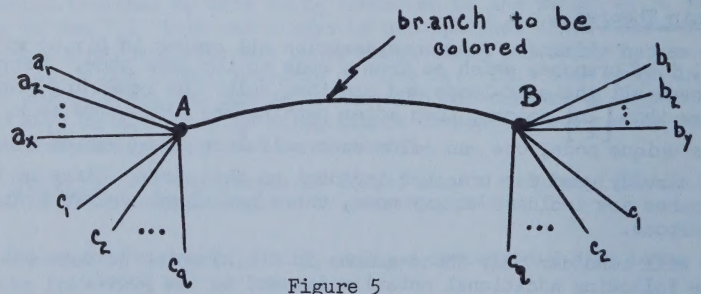


Figure 5

In this figure, $a_1, a_2 \ldots a_x$ (called the a colors); $b_1, b_2 \ldots b_y$ (called the b colors); $c_1, c_2 \ldots c_q$ (called the c colors) are each the color of the branch on which it appears, and no two of them are the same color.

Now we could easily color A - B, if we could change the color of any of the A branches colored a_i, i = 1 to x, to a b-color, without otherwise changing the set of colors of branches incident on A or B, because this would make the particular a_i thus changed available to color A - B. A change of this kind is prevented, however, if for each possible a_i, b_j combination with i = 1 to x, j = 1 to y, there is a path from A to B whose branches are alternately colored a_i and b_j. If there were such a path, then when we changed the color of the A-branch from a_i to b_j, we would be forced to change the next branch on this path to a_i, and then the next branch to b_j, and so on until the B-branch colored b_j at the end of the path was recolored a_i. The net effect is that the same set of colors still color the A and B branches, and still no color is available for A - B. On the other hand, if for some a_v,b_w such a path, of branches alternately colored a_v and b_w, did not exist, then there must be a path (path 1) starting at A consisting of branches alternately colored a_v and b_w which terminates on a node other than B (or A). By terminate on a node, we mean that the path in question enters that node with an a_v - branch in case 1, or a b_w - branch in case 2, but there is no b_w - branch in case 1, or a_v - branch in case 2 which is also incident on that node. Incidentally, we note there is a similar path (path 2) from B alternately colored b_w and a_j which terminates on a node other than A (or B). If we changed the color of all a_v - branches to b_w and all the b_w - branches to a_v on path 1, we would have a_v available to color A - B. If we made the same changes on path 2, we would have b_w available to color A - B.

In any case, we have shown thus far that either there is a path from A to B, alternately colored a_i and b_j for every combination of a_i, i = 1 to x, and b_j, j = 1 to y, or we can by some recoloring, find a way to legitimately color A - B. Let us assume from now on that all these alternately colored paths exist,or the graph has an "initial impasse".

Assumption A1, and the General Solution to Initial Impasse

We will now describe our general approach to coloring A - B in a graph with an initial impasse. To do this, we will have to assume some results which will be established later.

We can represent part of our graph as follows:

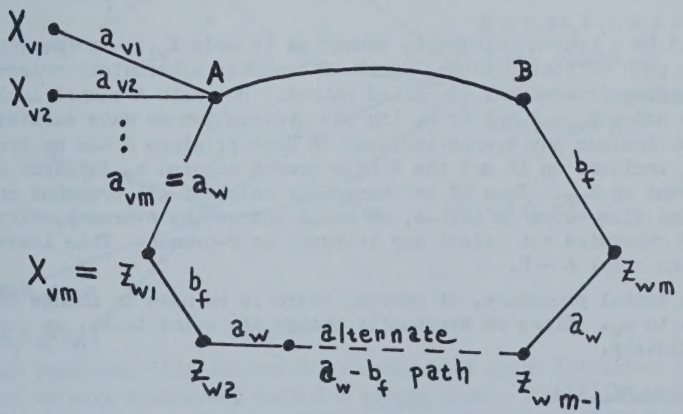


Figure 6

The situation pictured here is meant to imply the existence of a number of assumed conditions not evident from the picture alone:

(A1) We can find a sequence (called the sesame sequence) of m differently colored A - branches, colored $a_{v1}, a_{v2} \cdots a_{vm} (= a_w)$ respectively, incident on nodes labelled $X_{v1}, X_{v2} \cdots X_{vm}$ such that:

(a) the node labelled X_{v1} is distinct from the node X_{vm}, and

(b) the color of the branches colored $a_{v,j-1}$ may be changed legitimately to a_{vj}, for all j, j=2 to m, and the color of the A-branch colored a_{v1} can be changed to $a_{vm} = a_w$ without necessitating further changes.

(A1) is assumed to be true for now. It will be proven later.

Pictured also is a path from A to B whose branches are alternately colored a_w and b_f, whose existence follows from the graph having an initial impasse.

Concerning this path, we observe that no branch incident on Z_{wj}, j = 1 to m, other than $Z_{w,j-1} - Z_{wj}$, and $Z_{wj} - Z_{w,j+1}$ can be colored a_w or b_f.

Accepting the above picture and all it is meant to imply, we can describe the technique for coloring A - B.

Assume, as in (A1), that we can change the color of the A-branch colored a_{vj} to $a_{v,j-1}$ and the color of the A -branch colored a_{v1} to a_w. Now consider the subgraph consisting of branches colored a_w and b_f only, and the nodes on which such branches are incident. This subgraph includes the following:

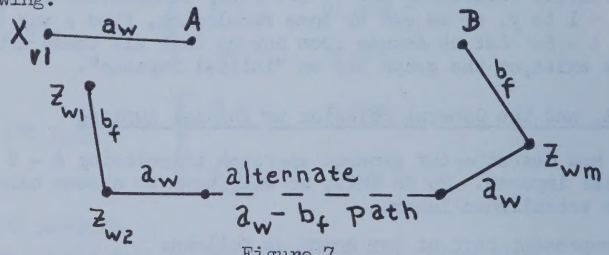


Figure 7

There could be a branch colored b_f connected to node X_{v1}. In fact, there could be a path of considerable length of branches alternately colored b_f and a_w emerging from node X_{v1}, called path-A. But path-A cannot include any of the nodes Z_{wi}, i = 1 to m, (by the observation we made earlier) and thus cannot include any branch incident on node B; since B has no branch colored a_w incident on it and the single branch colored b_f incident on B is also incident on Z_{wm}. Thus if we change the color of all branches colored a_w to b_f and vice-versa on path-A, we would change the A-branch colored a_w to b_f, and otherwise not effect any A-branch or B-branch. This leaves a_w available to color A - B.

In an actual procedure, of course, there is no need to change the color of $X_{v1} - A$ to a_w. Since we eventually change its color to b_f, we could do that immediately.

Justification of (A1)

Now we have the remaining problem of justifying assumption (A1)

Each A-branch colored with an a-color is also incident on another node. Let Y_i, i=1 to m be the nodes other than A upon which A-branches colored a are incident.

Then each node Y_i has also incident upon it y b-branches colored with b_j, j=1 to y. This follows from the initial impasse.

Thus if z is the number of a-branches (the number of different a colors of branches) incident on Y_i for any i=1 to m

$$z \leq n-y$$

z is maximum when y is minimum so using the minimum y given by (3b) we get

$$z \leq z_{max} = \begin{cases} \frac{n}{2} - 1 & \text{for n even} \\ \frac{n+1}{2} - 1 & \text{n odd} \end{cases}$$

On the other hand the minimum total number of a-branches is given by (3a) as

$$x_{min} \leq \begin{cases} \frac{n}{2} + 1 & \text{n even} \\ \frac{n-1}{2} + 1 & \text{n odd} \end{cases}$$

Therefore, subtracting z from x_{min}, for each node Y_i there is <u>at least one a-color</u> which does not color a branch incident on Y_i. (In particular if n is even there are at least two such a colors, and if n is odd there is a least 1). The definitions of Y_i node, and the conclusions concerning the m are summarized in Figure 8 below.

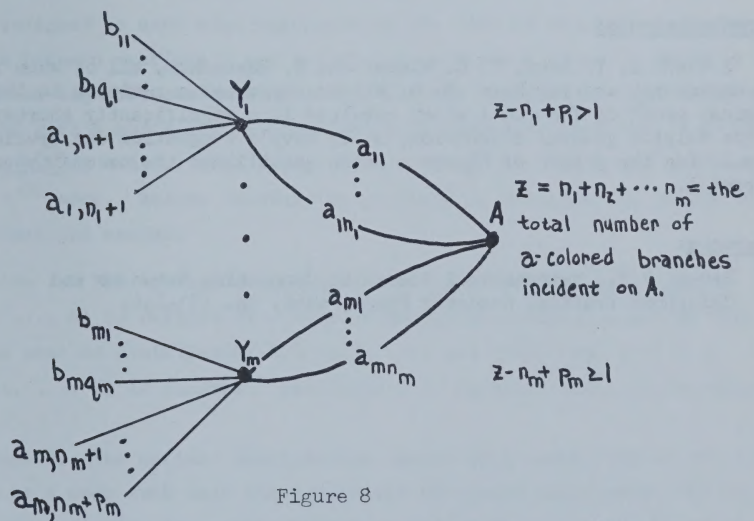


Figure 8

We can represent this situation by a directed graph (distinct from any other graph we have discussed) called a <u>sesame</u> graph. The sesame graph has m nodes, labelled 1,2,...,m, respectively corresponding to the m nodes labelled $Y_1, Y_2, \ldots Y_m$, of Figure 8. There is a directed branch from node

j to k ($j \neq k$) labelled a_x if the node Y_j has no branch labelled a_x incident on it, and Y_k has a branch labelled a_x incident upon it.

By the previous paragraphs there will be at least one directed branch leaving every node of the sesame graph. Therefore, there must be a simple directed circuit in the sesame graph. Corresponding to each such circuit of the sesame graph there is a sesame sequence. Let the path below be a circuit of a sesame graph.

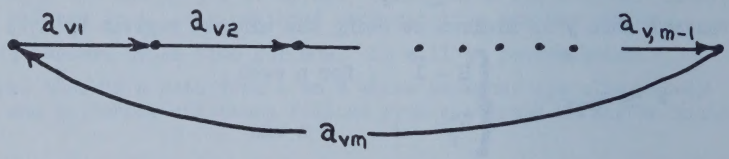


Figure 9

Then we know that we can change the A-Y_{p_1} branch colored a_{v1} to the color a_{v2}, and the only other branch colored a_{v2} will be one of the A - Y_{p_2} branches. This branch in turn can be recolored a_{v2}, and the only other branch which will be colored a_{v2} will be an A-Y_{p_3} branch, and so on until the only problem remaining is that we have two branches colored a_{vm} one of which is the A-Y_{p_0} branch which was originally colored a_{vm}. This branch may now be recolored a_{v1} and since the one A-branch, namely an A-$Y_{p_{v1}}$, which was formerly colored a_{v1} has been recolored with a_{v2} there is no conflict. Thus, the A-branches colored $a_{v1}, a_{v2}, \ldots a_{vm}$ are a sesame sequence.

Acknowledgements:

I thank S. Y. Levy, R. O. Winder and N. Maxemchuk, all of whom read the manuscript and further: R. O. Winder suggested a revision in the original proof of Theorem 1 which resulted in a significantly shorter proof. Beside helpful general discussion, S. Y. Levy's suggestion was crucial in formulating the graphs of Figure 4 which established the necessity part of Theorem 1.

Reference:

[1] Benes, V.E., Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press, 1965, pp. 153-146.

ON CERTAIN OPTIMIZATION PROBLEMS IN COMMUNICATION NETWORKS

S. Y. CHAN
University of California, Berkeley, California, and
San Diego State College, San Diego, California

W. S. CHOU
University of California, Berkeley, California, and
Network Analysis Corporation, Glen Cove, New York

ABSTRACT

The definitions of the optimal distribution tree with respect to a node and the optimal distribution center of a communication network are given. Based on dynamic programming techniques, this paper presents an effective step-by-step approach for determining both the optimal distribution tree with respect to a node and the optimal distribution center which have various engineering applications especially in the field of communication networks.

INTRODUCTION

For a variety of problems the nodes of a communication net can be considered as stations or distribution centers such that a positive number C_{ij} can be assigned to each edge representing the cost of transporting some commodities between stations i and j. If C_{ij} is given a value of infinity when there is no edge between nodes i and j, all the information concerning the given net with n nodes can be contained in a symmetric cost matrix $C = [C_{ij}]$ with $C_{ii} \triangleq (i)$ $(i = 1, \ldots, n)$ along the main diagonal indicating the position of the i^{th} node. Before stating the problems of interest the following definitions are needed:

Definition 1. The optimal distribution tree with respect to a node i $(i = 1, \ldots, n)$ is defined as a tree of the given communication net such that the cost of distribution from node i to any other node j $(j = 1, \ldots, i-1, i+1, \ldots, n)$ is minimum. This node i is defined as the distribution center.

Definition 2. The optimal distribution center of a communication net is defined as a node such that the sum of all the costs associated with the optimal distribution tree with respect to this node is not greater than the sum of all the costs associated with the optimal distribution tree with respect to any other node.

STATEMENT OF THE PROBLEM

With the above definitions, the problems of interest can be stated as follows:

(1) Given a communication net together with the cost matrix C, determine the optimal distribution tree with respect to a specified node.

(2) Given a communication net together with the cost matrix C, determine the optimal distribution center.

Different methods may be employed for determining the solutions of the problems posed above [1 - 4]. Since the problem of determining the path between any two nodes with minimal cost is essentially the same as that of determining the shortest path between two nodes, repeating the same process for finding the shortest path between two nodes n-1 times will result in the optimal distribution tree with respect to a specified node. However, by means of the method presented in this paper, all the optimal paths originated from a specified node can be determined simultaneously and hence the corresponding optimal distribution tree can be immediately obtained. A slight extension of this method yields the solution of the second problem.

METHOD OF SOLUTION

Let the given net consisting of n nodes be considered as a system with n distinct states $S \triangleq \{1, 2, \ldots, n\}$. Then the transition from one station to another can be regarded as the motion of the discrete system governed by a transformation of the form

$$x_{i+1} \triangleq T(x_i, u_i) \tag{1}$$

where $x_i \in S$ and $u_i \in I \triangleq \{1, 2, \ldots, n\}$ represent, respectively, the state and the input (control) of the system at the i^{th} stage. The transformation T in (1) is assumed to be a single-valued function of both x_i and u_i and is defined in the following manner: The state x_{i+1} in (1) assumes the value j if $u_i = j$ regardless of the value of the previous state x_i. Thus, for instance, if x_i assumes the value, say $x_i = k$, then $u_i = j$ demands that the system be transferred to the new state $x_{i+1} = j$. Of course, in order that any two successive states x_i and x_j represent two distinct states, u_i and x_i must assume different values. Let the cost for transporting from station x_1 to station x_{k+1} be defined by

$$L_k(x_1) \triangleq \sum_{i=1}^{k} L(x_i, u_i) \tag{2}$$

where $L(x_i, u_i) \triangleq C_{jk}$ for $x_i = j$ and $u_i = k$. Let the minimum value of $L_k(x_1)$ be denoted by $g_k(x_1)$, that is,

$$g_k(x_1) = \underset{\{u_i\}}{\text{Min}} \; L_k(x_1) = \underset{\{u_i\}}{\text{Min}} \left\{ \sum_{i=1}^{k} L(x_i, u_i) \right\}, \quad k = 1, 2, \ldots \tag{3}$$

Then, invoking the principle of optimality [5] $g_k(x_1)$ can be written as with $u_k = x_{k+1} = s$ and $s \in \{1, 2, \ldots, n\}$,

$$g_1(x_1) = L(x_1, s) \tag{4}$$

$$g_k(x_1) = \underset{u_1}{\text{Min}} \left\{ L(x_1, u_1) + g_{k-1}(x_2) \right\}, \quad k = 2, 3, \ldots \tag{5}$$

where $g_{k-1}(x_2)$ is the minimum of the sum of the last k-1 terms. Thus, in general, writing $g_k(x_j)$ explicitly, one finds

$$g_k(x_j) = \underset{\{u_i\}}{\text{Min}} \sum_{i=j}^{k+j-1} L(x_i, u_i) \tag{6}$$

If s is the distribution center to which the corresponding optimal distribution tree is to be found, the cost associated with the optimal path between s and an arbitrary station $r \in \{1, \ldots, s-1, s+1, \ldots, n\}$ is given by $\underset{k}{\text{Min}} \{g(r)\}$ where $k = 1, \ldots, n-1$. Let l_r be the integer such that

$$g_{l_r}(r) = \underset{k}{\text{Min}} \{g_k(r)\} \tag{7}$$

then the total cost associated with the optimal distribution tree with respect to node s is

$$T_s \triangleq \sum_{\substack{r=1 \\ r \neq s}}^{n} g_{l_r}(r) = \sum_{\substack{r=1 \\ r \neq s}}^{n} \underset{k}{\text{Min}} \{g_k(r)\}, \quad s = 1, 2, \ldots, n \tag{8}$$

and its determination will lead to the solution of problem 1 posed above. The solution of problem 2, that is, the determination of the optimal distribution center, is obtained by finding the minimum value of T_s in (8) in the following manner:

Let m_s be the integer such that

$$T_{m_s} = \underset{s}{\text{Min}}\ T_s, \quad s \in \{1, 2, \ldots, n\} \tag{9}$$

then m_s is the optimal distribution center.

Without loss of generality, let $s = 1$ and $x_1 = r$, then (4) and (5) become, respectively,

$$g_1(r) = L(r, 1) \qquad r = 2, 3, \ldots, n \tag{10}$$

and

$$\begin{aligned} g_k(r) &= \underset{u_1}{\text{Min}} \left\{L(r, u_1) + g_{k-1}(x_2)\right\} \\ &= \underset{u_1}{\text{Min}} \left[L(r, u_1) + g_{k-1}\left\{T(r, u_1)\right\}\right], \quad k = 2, \ldots, n \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad r = 2, 3, \ldots, n \end{aligned} \tag{11}$$

where, in the last expression of (11), the relation in (1) has been applied.

Consider the expressions in (10) and (11) from which, with $r = 2, 3, \ldots, n$, the following table can be easily established.

TABLE 1

r	2	3	. . .	n
u_{n-1}	-	-	. . .	-
g_1	-	-	. . .	-
u_{n-2}	-	-	. . .	-
g_2	-	-	. . .	-
.	.	.	. . .	.
.	.	.	. . .	.
u_2	-	-	. . .	-
g_{n-2}	-	-	. . .	-
u_1	-	-	. . .	-
g_{n-1}	-	-	. . .	-

For each of the n-1 columns in Table 1, one notes the smallest value of g_i ($i = 2, \ldots, n$) under column j ($j = 2, \ldots, n$), denoted by $g_{1_j}(j)$, i.e., for the j^{th} column,

$$g_{1_j}(j) \triangleq \underset{i}{\text{Min}} \left\{g_i\right\}, \quad j = 2, \ldots, n \tag{12}$$

With Table 1 together with $g_{1_j}(j)$ just defined, the solutions to problems (1) and (2) posed above can be determined as follows:

(I) Starting with g_{1_j}, one notes the value of u_{n-1_j}, say $u_{j-1_j} = a$, and then proceeds to column a to find the value of u_{n-1_j+1}, say $u_{n-1_j+1} = b$. Next, proceed to column b to determine the value of u_{n-1_j+2}. The next column to be investigated is, of course, dependent upon the value of u_{j-1_j+2} just found. Continuing this process until u_{n-1} is determined results in the 1_j-stage optimal policy

$$\underline{u}_j \triangleq (u_{n-1_j}, u_{n-1_j+1}, \ldots, u_{n-1}) \qquad j = 2, \ldots, n \tag{13}$$

which defines the optimal path from station j to station 1.

Adjoining all these optimal paths together yields the optimal distribution tree with respect to node 1. The total cost associated with this tree is given by (6) with s = 1, i.e.,

$$T_1 = \sum_{j=2}^{n} g_{1_j}(j) \tag{14}$$

With the above development, it is evident that the recurrence relation needed for the determination of the optimal distribution tree with respect to any node s can be obtained by modifying (10) and (11) slightly so that the following equations result

$$g_k(r) = \min_{u_1} \left[L(r, u_1) + g_{k-1} \left\{ T(r, u_1) \right\} \right], \quad k = 2, \ldots, n \tag{15}$$
$$r = 1, 2, \ldots, s-1, s+1, \ldots, n$$

and

$$g_1(r) = L(r, s) \qquad r = 1, 2, \ldots, s-1, s+1, \ldots, n \tag{16}$$
$$s = 1, 2, \ldots, n$$

By means of (15) and (16) a table similar to Table 1 can be set up for each value of s, thus enabling one to determine the optimal distribution tree with respect to any node s by the same procedure as outlined above.

(II) To determine the optimal distribution center, one needs to determine the optimal distribution tree with respect to each of the nodes by the

procedure outlined in (I). The optimal distribution center is the node associated with the particular optimal distribution tree such that the total cost is minimal, i.e., $\underset{i}{\text{Min}} \{T_i\}$. Thus, for instance, if T_a is the smallest of all T_i's, viz.,

$$T_a = \underset{i}{\text{Min}} \{T_i\} \tag{17}$$

then node a is the optimal distribution center of the given net.

REFERENCES

1. K. Steiglitz, P. Weiner, and D. Kleitman, "The Design of Minimum Cost Survivable Networks," to appear in IEEE Transactions on Circuit Theory.

2. S. Lin, "Computer Solutions of the Traveling Salesman Problems," Bell Sys. Tech. Journal, Vol. 44, No. 10, pp. 2245-2269, December 1965.

3. R. Bellman, "Dynamic Programming Treatment of the Traveling Salesman Problem," J. ACM, 61-63, 1962.

4. K. Steiglitz, P. Weiner, "Some Improved Algorithms for Computer Solution of the Traveling Salesman Problem," Proc. Six Annual Allerton Conference on Circuit and System Theory, pp. 814-821, 1968.

5. J. T. Tou, Modern Control System, McGraw-Hill, 1964.

FAULT DIAGNOSIS OF LINEAR SYSTEMS*

M. F-S LIN and S. PARK CHAN
Department of Electrical Engineering
University of Santa Clara, Santa Clara, California

ABSTRACT

A large system may be considered as an interconnection of a number of subnetworks which are comprised of various combinations of linear and nonlinear circuits. It is shown that by using the relationships between the over-all voltage gain of the system and the voltage gains of its subnetworks at a selected set of test frequencies, a table, called the one-component-variation table, is obtained.

A fault diagnosis procedure is developed for a linear system of cascade connected two-port networks based on the one-component-variation table and the measurements of the magnitudes of the voltage gains made at the selected frequencies.

The fault testing procedure discussed in this paper is based on a theorem (Theorem 1) which indicates that the voltage gain G_j of subnetwork N_j of a given network N is independent of the voltage gain of the subnetworks N_1, N_2, ... , N_{j-1} preceding N_j. Using Eq. (7) in the theorem, along with the one-component-variation table one can detect the subnetwork which contains the faulty component.

An example is worked out in detail to demonstrate the fault diagnosis procedure.

INTRODUCTION

Recently the problem of fault diagnosis in linear systems has been studied by a number of investigators [1], [2] and [4]. In this paper, a fault diagnosis technique for linear systems is developed which requires the generation of a "one-component-variation table" and the evaluation of its voltage gain and the voltage gains of its cascade connected subnetworks at a selected set of test frequencies. It is shown that the value of the voltage gain in a subnetwork is dependent only upon its component values and the component values of the subsequent subnetworks.

---

* This work is supported in part by the National Aeronautics and Space Administration, Grant No. NGR-05-017-017.

Using the theory of cascade connected two-ports, identification of the faulty components in a given system can be achieved by comparing the external measurements of the system with the information contained in the table.

THEORETICAL DEVELOPMENT

The relationships between the over-all voltage gain of the system and the voltage gains of its subnetworks constitute the most important information to be obtained in the procedure of fault diagnosis. The following theorem may be stated:

Theorem 1: Given a linear two-port N consisting of n cascade connected two-port subnetworks $N_1, N_2 \ldots, N_n$ as shown in Figure 1. Let G_j be the voltage gain of the j^{th} subnetwork N_j. Then, G_j is independent of the voltage gains $G_1, G_2, \ldots, G_{j-1}$ of the subnetworks $N_1, N_2, \ldots, N_{j-1}$ which precede N_j.

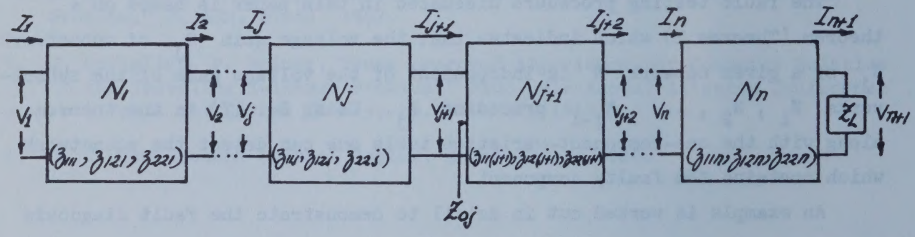


Figure 1

Proof: Referring to Figure 1, we write

$$V_j = z_{11j} I_j + z_{12j} (-I_{j+1}) \tag{1}$$

$$V_{j+1} = z_{21j} I_j + z_{22j} (-I_{j+1}) \tag{2}$$

Since $V_{j+1} = I_{j+1} Z_{oj}$, we have

$$I_{j+1} = \frac{V_{j+1}}{Z_{oj}} \tag{3}$$

Substituting (3) into (1) and (2), we have

$$V_j = z_{11j} I_j - \frac{z_{12j}}{Z_{oj}} V_{j+1} \tag{4}$$

$$V_{j+1} = z_{21j} I_j - \frac{z_{22j}}{Z_{oj}} V_{j+1} \tag{5}$$

From (4)

$$I_j = \frac{V_j}{z_{11j}} + \frac{z_{12j}}{Z_{oj} z_{11j}} V_{j+1} \tag{6}$$

Substituting (6) into (5), we get

$$V_{j+1} = \frac{z_{21j}}{z_{11j}} V_j + \frac{z_{12j} z_{21j}}{Z_{oj} z_{11j}} V_{j+1} - \frac{z_{22j}}{Z_{oj}} V_{j+1}$$

from which it can be shown that

$$G_j = \frac{V_{j+1}}{V_j} = \frac{z_{21j} Z_{oj}}{z_{11j} z_{22j} - z_{12j} z_{21j} + z_{11j} Z_{oj}} \tag{7}$$

where Z_{oj} is the output impedance of N_j and z_{11j}, z_{12j}, z_{21j}, z_{22j} are the open-circuit impedances of N_j [3].

From (7), it is evident that G_j is independent of $G_1, G_2, \cdots, G_{j-1}$. The fault diagnosis procedure may now be stated as follows:

Fault Testing Procedure

Step (1). Partition the given system into a number of cascade connected two-port subnetworks $N_1, N_2, \cdots, N_n$.

Step (2). Select a set of test frequencies (including the critical frequencies which may be obtained from the nominal gain expression or frequency response curve).

Step (3). Compute the nominal values in db of the voltage gains $G_1, G_2, \cdots, G_n$, and G at the selected test frequencies (where G is the over-all voltage gain of N).

Step (4). Vary the value of each component of the given system above and below its nominal value.

Step (5). Calculate the voltage gains in db at each of the test frequencies for each deviant value of each component in the given system.

Step (6). A "gain signature" is used to indicate the degree of deviation in db of a voltage gain from its nominal value at a test frequency. Thus, with each component variation is associated a gain signature, which, in coded form, gives the corresponding gain deviation at a test frequency.

Step (7). Three types of frequency responses are used which are defined as follows:

Type 1: The symbol P^- is used to indicate that the voltage gain in db of the system N (or a subnetwork N_j of N) is decreasing

with the increasing values of the test frequencies.

Type 2: The symbol P^+ is used to denote the fact that the voltage gain in db of N (or N_j) is increasing with the increasing values of the test frequencies.

Type 3: The symbol Q is used to indicate that the voltage gains in db of N (or N_j) increases at one frequency value and decreases at another with the increasing values of the test frequencies.

Step (8). Finally, combining the information obtained in Steps (6) and (7), a table, called one-component-variation table, is formulated which gives the necessary data for fault identification.

For example, for a given network N (or N_j), a particular composite symbol, say P^- (5576), for the voltage gain G of N (or G_j of N_j) with respect to a deviant value of a component (or a set of component values) indicates that G (or G_j) is decreasing in db with increasing test frequency values and that the db values of the voltage gain vary according to its gain signatures 5, 5, 7, and 6 (to be defined in each problem, see the example in the following section) at the test frequencies f_1, f_2, f_3 and f_4, respectively.

EXAMPLE

Example. Consider the equivalent circuit of a one-stage transistor amplifier as shown in Figure 2. The fault diagnosis of this network is carried out, according to the procedure outlined in Section II, in the following steps:

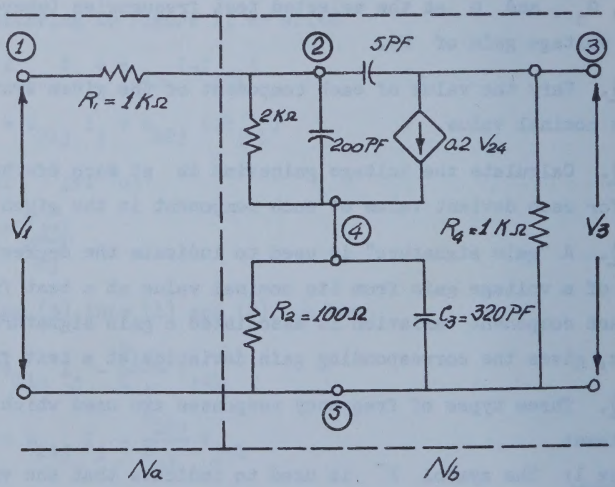


Figure 2

Step (1). Partition the network N into two subnetworks N_a and N_b as shown in Figure 2.

Step (2). Select a set of four test frequencies f_1, f_2, f_3 and f_4 as follows:

$f_1 = 4 \times 10^4$ Hz, $f_2 = 4 \times 10^5$ Hz,
$f_3 = 4 \times 10^6$ Hz, $f_4 = 1 \times 10^7$ Hz.

Step (3). Compute the nominal values of the voltage gains G, G_a, and G_b as shown in Table II.

Step (4). Change the value of each of the components R_1, R_2, C_3, R_4 as shown in Table I.

Component	Nominal Value	Above(↗) Nominal Value	Below(↘) Nominal Value
R_1	1000 Ω	100 kΩ	10 Ω
R_2	100 Ω	10 kΩ	1 Ω
C_3	320 PF	32000 PF	3.2 PF
R_4	1000 Ω	100 kΩ	10 Ω

TABLE I

Step (5). Calculate the voltage gains (db) at each of the test frequencies for each deviant value (Table I) of each component in each of the networks N, N_a, and N_b as shown in Table II.

F	$F_1 = 4 \times 10^4$ HZ			$F_2 = 4 \times 10^5$ HZ		
db / Comp.	G	G_a	G_b	G	G_a	G_b
Nominal Value	19.351	-0.204	19.555	19.35	-0.236	19.587
R_1 ↗	8.151	-11.403	19.555	-5.246	-24.833	19.587
R_1 ↘	19.553	-0.002	19.555	19.585	-0.001	19.587
R_2 ↗	-17.914	-0.001	-17.912	-1.081	0.121	-1.202
R_2 ↘	41.271	-3.161	44.433	34.54	-9.891	44.432
C_3 ↗	21.593	-0.121	21.715	39.478	1.674	37.803
C_3 ↘	19.35	-0.205	19.555	19.212	-0.341	19.554
R_4 ↗	55.101	-4.386	59.487	36.335	-18.633	54.968
R_4 ↘	-20.647	-0.203	-20.444	-20.604	-0.192	-20.412

TABLE II

F	$F_3 = 4 \times 10^6$ HZ			$F_4 = 10^7$ HZ		
db / Comp.	G	G_a	G_b	G	G_a	G_b
Nominal Value	16.757	-4.847	21.604	8.818	-17.145	25.963
R_1 ↗	-24.144	-45.749	21.604	-31.84	-57.804	25.963
R_1 ↘	21.678	0.073	21.604	26.372	0.408	25.963
R_2 ↗	25.782	7.872	17.909	9.389	-15.991	25.381
R_2 ↘	16.269	-28.094	44.364	8.364	-35.658	44.023
C_3 ↗	16.354	-29.261	45.615	8.388	-37.121	45.509
C_3 ↘	13.63	-5.856	19.487	6.831	-12.317	19.148
R_4 ↗	17.335	-22.274	39.610	9.684	-26.74	36.425
R_4 ↘	-17.71	0.615	-18.326	-14.954	-1.327	-13.627

TABLE II

Step (6). The quantization of the "gain signatures" is done according to the following conventions:

(0) within 1 db of nominal
(1) 1 to 10 db too high
(2) 10 to 20 db too high
(3) 20 to 30 db too high
(4) more than 30 db too high
(5) 1 to 10 db too low
(6) 10 to 20 db too low
(7) 20 to 30 db too low
(8) more than 30 db too low

Step (7) and (8). A one-component-variation Table is formulated as shown in Table III.

	G	G_a	G_b
Nominal	$P^-(0000)$	$P^-(0000)$	$P^+(0000)$
R_1 ↗	$P^-(6788)$	$P^-(6788)$	$P^+(0000)$
R_1 ↘	$P^+(0012)$	$P^+(0012)$	$P^+(0000)$
R_2 ↗	Q(8710)	Q(0025)	$P^+(8750)$
R_2 ↘	$P^-(3200)$	$P^-(5576)$	$P^-(3332)$
C_3 ↗	Q(1300)	Q(0176)	Q(1232)
C_3 ↘	$P^-(0055)$	$P^-(0051)$	$P^-(0055)$
R_4 ↗	$P^-(4200)$	$P^-(5665)$	$P^-(4422)$
R_4 ↘	$P^+(8887)$	Q(0012)	$P^+(8888)$

TABLE III

Two different cases of component faults are now discussed.

Case 1. Let the external measurements of the given network N (Fig. 2) be represented in composite symbolic forms as follows:

G	G_a	G_b
$P^-(6788)$	$P^-(6788)$	$P^+(0000)$

Comparing the above information with that in Table III, it is seen that the signatures of G_b are unchanged, and in light of Theorem 1, the fault must occur in N_a. Next, comparing the measured signatures of G_a and G with those listed in Table III, it is evident that the faulty component is R_1 with its actual value higher than its nominal value.

Case 2. Let the external measurements of N be given as follows:

G	G_a	G_b
$P^-(5788)$	$P^-(6888)$	$Q(1232)$

Since the measured values of G_b are different from its nominal values at the four test frequencies, the fault must occur in both N_a and N_b according to Theorem 1. By comparing the measured signatures with those in Table III, the faulty components are readily identified as R_1 and C_3, both being above their nominal values.

CONCLUSIONS

A fault diagnosis technique has been developed in the above discussions. Although the example used here is quite simple, the fault diagnosis precedure presented in this work is a general one and can be applied to any linear system.

The main differences in approach for fault diagnosis between this work and the works cited in the References are as follows:

(1) In the fault testing procedure developed in this paper, fault detection is done by identifying the faulty "section" or subnetwork using Theorem 1 and then pin-pointing the faulty element from the subnetwork whereas in all other works cited, fault identification was accomplished by considering the given network as a whole. Because of this, one can handle a system with high complexity using the present method.

(2) From the one-component-variation table, one can detect not only faults of the one-element types but also multi-element faults in the given system using the same (one-component-variation) table, as illustrated in Case 2 of the example (on this page). That is, although References [1] and [2] also deal with multi-element faults, the techniques used in both papers require the generation of an extensive table listing all possible element

combinations, whereas in the method developed in this work only a simple one-component-variation table is necessary. Hence, fault identification can be accomplished with higher speed.

In summary, the main contribution of this work is the presentation of a simpler fault identification technique which can be applied to a more complex system with higher speed and efficiency than those techniques cited in the References.

Although it has been assumed that the system is comprised of a number of cascade connected subnetworks, this assumption of cascade connections among the subnetworks can be removed by redefining any voltage gain, say G_j, as a ratio of certain two node voltages of the given system with respect to a chosen reference node, so that the one-component-variation table will now contain the specified node-voltage ratios calculated at a selected set of nodes, and the comparison is made with the measured values against the corresponding calculated values at the test frequencies.

Finally, for a large system consisting of many subnetworks, the generation of the one-component-variation table will be time consuming, and hence should be done using certain automated procedures.[5]

REFERENCES

[1] S. Seshu and R. Waxman, "Fault Isolation in Conventional Linear Systems-A Feasibility Study", IEEE Trans. on Reliability, Vol. R-15, No. 1, pp. 11-16, May, 1966.

[2] W. J. Stahl, J. H. Maenpaa, C. J. Stehman, "Fault Diagnosis by Computer," Electronics, January 22, 1968, pp. 64-67.

[3] S. P. Chan, "Introductory Topological Analysis of Electrical Networks," Holt, Rinehart and Winston, 1969.

[4] W. J. Stahl, J. H. Maenpaa and C. J. Stehman, "Development of Advanced Dynamic Fault Diagnosis Techniques," Tech. Rep., AFAPL-TR-67-44, May, 1967.

[5] F. Rode and S. P. Chan, "Evaluation of Topological Formulas Using Digital Computers," Electronics, Vol. 4, No. 12, pp. 256-258, June, 1968.

STATE ESTIMATION, SYSTEM IDENTIFICATION, AND SIGNAL DETECTION

KEVIN A. CLEMENTS
Singer-General Precision, Inc.
Little Falls, New Jersey

ABSTRACT

Certain problems in state estimation, system identification, and signal detection are special cases of the problem which is considered in this paper. The class of systems considered is described by the vector stochastic differential equation, $dx(t) = f[x(t), t, \alpha]dt + \sigma[x(t), t, \alpha]dz(t)$, in which $z(t)$ is a Wiener process. The random variable (α) is discretely distributed. Observations of the form, $dy(t) = h[x(t), t, \alpha]dt + R^{1/2}(t)\, dw(t)$ where $w(t)$ is a Wiener process, are made. The apriori density function of the initial state, $x(t_o)$, and the apriori probabilities governing (α) are specified. A set of stochastic differential equations which describe the temporal evolution of the probabilities governing (α), is derived. These results, together with some previous results, enable one to completely specify the a posteriori density function of the state which solves the state estimation problem. Applications of these results to specific problems are given.

INTRODUCTION

The problem of state estimation for systems described by nonlinear stochastic differential equations with an unknown parameter, discretely distributed, is studied in this paper. Exact dynamic equations governing the temporal evolution of the conditional probability density of the state are obtained, and from these, the conditional expectation can be represented as the solution of a set of stochastic differential equations.

Several authors have considered the exact solution to the nonlinear filtering problem. Stratonovich (1)* was the first to do this, however, his results are not consistent with the Ito stochastic calculus due to the neglect of certain second-order terms. Kushner (2,3) corrected this error and formally derived the correct expression for the conditional density and later (4) rigorously derived some of these results. Bucy (5,6) has provided a different and somewhat simpler approach to the problem with the representation theorem. It should be noted that the state estimation problem for systems with unknown parameters, continuously distributed, is included in the general nonlinear estimation problem considered by these authors; however, if the parameters are discretely distributed, this is not true. The results presented in this paper represent a direct extension of the nonlinear filtering theory. To provide a representation for the conditional density and the optimum estimate, the results for the nonlinear case (which have been previously derived) are used together with the stochastic differential equations for the conditional probabilities of the unknown parameters obtained in this paper. Magill (7) has done some parallel work for the linear gaussian discrete time case.

In addition to providing a solution to the problems of adaptive estimation and system identification, the results of this paper can be applied to the problem of detecting random or deterministic signals in noise.

*Numbers in brackets (-) refer to references at the end of this paper.

The results are specialized to the linear gaussian case in which the optimum estimate is generated by a set of Kalman Filters whose outputs are weighted by the conditional probabilities for the parameter.

PROBLEM STATEMENT

The system under consideration is described by the (n) dimensional vector stochastic differential equation:

$$dx(t) = f[x(t),t,\alpha\,]dt + \sigma\,[x(t),t,\alpha\,]\,dz\,(t), \qquad [1]$$

where (d) denotes a differential; i.e. $dx(t)=x(t+dt) - x(t)$ and $z(t)$ is an (r) dimensional vector valued Wiener process.

The (m) dimensional measurement vector is described by:

$$dy(t) = h[x(t),t,\alpha\,]\,dt + R^{1/2}(t)\,dw(t), \qquad [2]$$

where $w(t)$ is an (m) dimensional Wiener process. It is assumed that $z(t)$ and $w(t)$ are independent, although a derivation can be carried out without this restriction. The unknown parameter (α) is a discrete random variable with range of values $k_1 \text{---} k_q$ and a priori probability distribution $\{p_i(o), i=1, \text{---} q\}$ at $t = 0$.

The functions $f[x(t),t,\alpha\,]$ and $\sigma\,[x(t),t,\alpha\,]$, are assumed to be Lipschitz in $x(t)$, and $R(s)$ is assumed positive definite and continuous at each s, $s\epsilon[t_o,t]$. $g[x(t)]$ is an arbitrary function satisfying satisfying suitable conditions to be specified. Let O_t be the minimal σ - field determined by the process $y(s)$, $s\epsilon[t_o,t]$. The problem can now be precisely stated; find the dynamical equations satisfied by the conditional density $p[x(t),t\,|\,O_t]$ of $x(t)$ and the conditional expectation $E\{g[x(t)]\,|\,O_t\}$ conditioned on O_t.

PRELIMINARIES

Some definitions and known results in nonlinear filtering theory are presented in this section. The reader is referred to (1-6) for a detailed account.

The conditional probability $p_i(t)$ is defined by

$$p_i(t) = P\{\alpha = k_i \,|\, O_t\}\,.$$

One can now write the conditional density $p[x,t \,|\, O_t\,]$ as

$$p[x,t \,|\, O_t\,] = \sum_{i=1}^{q} p_i(t)\; p[x,t \,|\, O_t,\; \alpha = k_i\,]\,. \qquad [3]$$

Bucy, (5-6), has derived an expression for the conditional density for the case in which (α) is known. This "representation theorem", with the notation suitably modified is repeated here without proof.

THEOREM 1. (BUCY)

Suppose there exists on the interval $[t_o, t\,]$ a unique continuous sample function Markov process $x(t)$, a solution of [1] for (α) known, having all its joint probability distributions

absolutely continuous with respect to Lebesque measure. Further assume that:

$$E \exp\left\{ \sup_{s \in [t_o, t]} \| h[x(s), s, k_i] \|^2_{R^{-1}(s)} (t - t_o) \right\} < \infty \qquad i = 1, \text{---} q.$$

Then the conditional density of $x(t)$ given O_t and $\alpha = k_i$ has a density:

$$p[x, t \mid O_t, \alpha = k_i] = \frac{E\left\{ e^{\Theta_i(t)} \mid O_t, \alpha = k_i, \bar{x}(t) = x \right\} \mu[x, t \mid \alpha = k_i]}{E\left\{ e^{\Theta_i(t)} \mid O_t, \alpha = k_i \right\}}, \qquad [4]$$

where $\mu[x, t \mid \alpha = k_i]$ is the density of $x(t)$ conditioned by $\alpha = k_i$. Let $z(t)$ and $\bar{z}(t)$ be independent Wiener processes, then $\bar{x}(t)$ is defined to be a solution of Eq. 4 in which $dz(t)$ is replaced by $d\bar{z}(t)$. $\Theta_i(t)$ is defined by,

$$\Theta_i(t) = -\frac{1}{2} \int_{t_o}^{t} \| h[\bar{x}(s), s, k_i] \|^2_{R^{-1}(s)} ds + \int_{t_o}^{t} h'[\bar{x}(s), s, k_i] R^{-1}(s) \, dy(s), \qquad [5]$$

with the second integral on the right of Eq. 5 a stochastic Ito integral.

COROLLARY 1

Under the assumptions of Theorem 1 and with:

$$E\left\{ | g[x(t)] | \right\} < \infty,$$

then:

$$E\left\{ g[x(t)] \mid O_t, \alpha = k_i \right\} = \frac{E\{ g[\bar{x}(t)] e^{\Theta_i(t)} \mid O_t, \alpha = k_i \}}{E\{ e^{\Theta_i(t)} \mid O_t, \alpha = k_i \}} \qquad [6]$$

Corrollary 1 follows from Theorem 1 by directly evaluating the expected value.

The following theorem, due to Ito [8] enables one to determine the stochastic differential equation satisfied by a function of a Markov process.

THEOREM 2 (ITO)

Let $x(t)$ be the unique solution of Eq. 1, with $\alpha = k_i$; then $\varphi[x(t), t]$, where φ is continuously differentiable in t and twice continuously differentiable in x, satisfies:

$$d\varphi[x(t), t] = \varphi_t[x(t), t] dt + \sum_i \varphi_{x_i}[x(t), t] d x_i(t) + \frac{1}{2} \sum_{ik} \varphi_{x_i x_k}[x(t), t] S_{ik}[x(t), t, k_i] dt, \qquad [7]$$

where $S_{ik}[x(t), t, k_i]$ is the i, kth element of the matrix $S[x(t), t, k_i] = \sigma[x(t), t, k_i] \cdot \sigma'[x(t), t, k_i]$.

COROLLARY 2

Let $\psi[x(t), t]$ be the scalar function, $\psi[x(t), t] = \gamma[x(t), t] / \delta[x(t), t]$, then

$$d\psi[x(t), t] = \left\{ \delta[x(t), t] d\gamma[x(t), t] - \gamma[x(t), t] d\delta[x(t), t] \right\} / \delta^2[x(t), t] +$$

$$\sum_{ik} \left\{ \gamma\,[x(t),t]\,\delta_{x_i}\,[x(t),t]\,\delta_{x_k}\,[x(t),t] - \delta[x(t),t]\,\gamma_{x_i}\,[x(t),t]\,\delta_{x_k}\,[x(t),t] \right\}$$

$$\cdot S_{ik}\,[x(t),t,k_j]/\delta^3\,[x(t),t]. \qquad [8]$$

Corollary 2 follows directly from Ito's theorem.

The density $\mu\,[x,t \mid \alpha = k_j]$ satisfies Kolmogorov's forward equation which, in differential form, is given by:

$$d\mu[x,t \mid \alpha = k_j] = \tilde{A}\mu\,[x,t \mid \alpha = k_j]\,dt,$$

where

$$\tilde{A}\cdot = -\sum_i \frac{\partial}{\partial x_i}\left\{ f_i\,[x(t),t,k_j]\cdot \right\} + \frac{1}{2}\sum_{i,k}\frac{\partial^2}{\partial x_i\,\partial x_k}\left\{ S_{ik}[x(t),t,k_j]\cdot \right\}$$

An equation describing the temporal evolution of $p[x,t \mid O_t, \alpha = k_j]$ can be obtained by applying Corollary 2 together with the Kolmogorov forward equation. It is given by:

$$dp[x,t \mid O_t, \alpha = k_j] = \tilde{A}p[x,t \mid O_t, \alpha = k_j] + \left\{ h\,[x(t),t,k_j] - \hat{h}_j \right\}' R^{-1}(t)\left\{ dy(t) - \hat{h}_j\,dt \right\}, \qquad [9a]$$

where $\hat{h}_j = E\left\{ h\,[x(t),t,\alpha] \mid O_t, \alpha = k_j \right\}$.

Applying Corollary 2 to Eq. 6 results in the equation:

$$d\hat{g}_j = (\widehat{Ag})_j\,dt + \left\{ (\widehat{gh})_j - \hat{g}_j\hat{h}_j \right\}' R^{-1}(t)\left\{ dy(t) - \hat{h}_j\,dt \right\}, \qquad [9b]$$

where:

$$A\cdot = \sum_i f_i\,[x(t),t,k_j]\frac{\partial}{\partial x_i}(\cdot) + \frac{1}{2}\sum_{ik} S_{ik}\,[x(t),t,k_j]\frac{\partial^2}{\partial x_i\,\partial x_k}(\cdot)$$

($\tilde{A}$ is the formal adjoint of A),

$$(\widehat{Ag})_j = E\left\{ Ag\,[x(t)] \mid O_t, \alpha = k_j \right\},$$

and

$$(\widehat{gh})_j = E\left\{ g[x(t)]\,h\,[x(t),t,\alpha] \mid O_t, \alpha = k_j \right\}.$$

Eq. 9b was proved rigorously by Kushner (4). Either Eq. 9a or Eq. 9b represents a solution to the nonlinear filtering problem when $\alpha = k_j$. Even for this case, exact computation is impossible in general, and approximations must be used to obtain a solution. For the linear case (dynamics and measurement), with a gaussian initial density for x(t), these results reduce to the Kalman Filter equations.

SUMMARY OF SIGNIFICANT RESULTS

The major results of this paper can be summarized in a theorem and a corollary which are proved in the appendix. Theorem 1 provides an expression for the conditional probability $p_j(t)$; Ito's theorem is then applied to this expression to obtain a stochastic differential equation for $p_j(t)$.

THEOREM 3

Under the conditions of Theorem 1, the probability that the random variable (α) is equal to p_i, conditioned on O_t, is given by:

$$p_i(t) = \frac{p_i(0) E\left\{e^{\Theta_i(t)} \mid O_t, \alpha = k_i\right\}}{\sum_{k=1}^{q} p_k(0)\, E\left\{e^{\Theta_k(t)} \mid O_t, \alpha = k_i\right\}}, \quad i = 1, \ldots q. \qquad [10]$$

Theorem 3 provides an expression analogous to the representation theorem which is given by Eq. 4. Application of Eq. 8 to this expression provides a stochastic differential equation for $p_i(t)$ which is stated in Corollary 3.

COROLLARY 3

The stochastic differential equation governing $p_i(t)$ is given by:

$$dp_i(t) = p_i(t)\left\{\hat{h}_i - \hat{h}\right\} R^{-1}(t)\left\{dy(t) - \hat{h}dt\right\}, \qquad [11]$$

where;

$$\hat{h} = \sum_{k=1}^{q} p_k(t)\, \hat{h}_i.$$

Eqs. 3,9 and 11 describe the temporal evolution for the conditional density and conditional probabilities of interest. Knowledge of these enables one to obtain the estimate of the state as well as to compute the probabilities that the system from which the measurements were obtained had parameter $\alpha = k_i$. This effectively gives an exact solution for both the state estimation and system identification problems. For the case of no dynamics and no process noise in which x is a scalar assuming values $k_1 \ldots k_i$ and h (x)=x, Eq. 11 reduces to an expression which was derived by Wonham [9]. For the discrete time case in which the measurement is given by: $y(i) = h[x(i), i, \alpha] + v(i)$, where $v(i)$ is gaussian with:

$$E\left\{v(i)\, v(j)'\right\} = R(i)\, \delta_{ij}.$$

a difference equation corresponding to Eq. 11 is obtained by applying Baye's rule recursively. It is given by:

$$p_i(i) = \frac{p_i(i-1)\, p\left\{y(i) \mid O_{i-1}, \alpha = k_i\right\}}{\sum_{k=1}^{q} p_k(i-1)\, p\left\{y(i) \mid O_{i-1}, \alpha = k_k\right\}} \qquad [12]$$

where O_i is the minimal σ - field determined by the measurement sequence $\left\{y(j)\ 1 \leq j \leq i\right\}$.

EXAMPLE 1

Suppose that the system dynamics and the measurement are linear, i.e.

$$\left.\begin{aligned} dx(t) &= F(t,\alpha)\, x(t)\, dt + G(t,\alpha)\, dz(t), \\ dy(t) &= H(t,\alpha)\, x(t)\, dt + R^{1/2}(t) dw(t). \end{aligned}\right\} [13]$$

In addition, let the initial densities $p[x, t_o \mid \alpha = k_i]$ be gaussian. It is well known that $p[x, t \mid O_t, \alpha = k_i]$ will be gaussian with mean $\hat{x}_i(t)$ and covariance $P_i(t)$ which are determined from the Kalman Filter equations, in which $\alpha = k_i$ is a known constant. The equation for $p_i(t)$ in this case is given by:

$$dp_i(t) = p_i(t)\left\{H(t,k_i)\,\hat{x}_i(t) - \widehat{H(t,\alpha)x(t)}\right\}' R^{-1}(t)\left\{dy(t) - \widehat{H(t,\alpha)x(t)}\,dt\right\}, \qquad [14]$$

where:

$$\widehat{H(t,\alpha)\,x(t)} = \sum_{i=1}^{q} p_i(t) H(t,k_i)\hat{x}_i(t). \qquad [15]$$

The optimum state estimate is given by:

$$\hat{x}(t) = \sum_{i=1}^{q} p_i(t)\,\hat{x}_i(t), \qquad [16]$$

and the covariance matrix of the error in the estimate is:

$$P(t) = \sum_{i=1}^{q} p_i(t)\left\{P_i(t) + [\hat{x}(t) - \hat{x}_i(t)]\,[\hat{x}(t) - \hat{x}_i(t)]'\right\} \qquad [17]$$

Eqs. 14, 15 and 16 follow directly from Eq. 3.

For the discrete time case, Eq. 12 is replaced by

$$\begin{aligned} x(i) &= \Phi(i \mid i-1, \alpha)x(i-1) + G(i,\alpha)\,u(i), \\ y(i) &= H(i,\alpha)\,x(i) + R^{1/2}(i)\,v(i). \end{aligned} \qquad [18]$$

where $u(i)$ and $v(i)$ are sequences of independent (zero mean) vector gaussian random variables with covariance matrices equal to the identity matrix. The initial densities $p[x, o \mid \alpha = k_i]$ are again assumed gaussian. In this case, the density $p[x, i \mid O_i, \alpha = k_i]$ is gaussian with mean $x_i(i)$ and covariance $P_i(i)$ determined by the discrete time Kalman Filter equations. The recursive equation for $p_i(i)$ is obtained from Eq. 12. It is given by:

$$p_i(i) = \frac{p_i(i-1)\,|\,M_i(i \mid i-1)|^{-1/2} \exp -\frac{1}{2}[\tilde{y}_i(i \mid i-1)'\,M_i^{-1}(i \mid i-1)\tilde{y}_i(i \mid i-1)]}{\sum_{k=1}^{q} p_k(i-1)\,|\,M_k(i \mid i-1)|^{-1/2} \exp -\frac{1}{2}[\tilde{y}_k(i \mid i-1)'\,M_k^{-1}(i \mid i-1)\tilde{y}_k(i \mid i-1)]}, \qquad [19]$$

where;

$$\begin{aligned} \tilde{y}_i(i \mid i-1) &= y(i) - H(i,k_i)\hat{x}_i(i \mid i-1), \\ M_i(i \mid i-1) &= H(i,k_i)P_i(i \mid i-1)\,H'(i,k_i) + R(i), \\ \hat{x}_i(i \mid i-1) &= E\{x(i) \mid O_{i-1}, \alpha = k_i\}, \\ P_i(i \mid i-1) &= E\{[x(i) - \hat{x}_i(i \mid i-1)]\,[x(i) - \hat{x}_i(i \mid i-1)]' \mid O_{i-1}, \alpha = k_i\}. \end{aligned}$$

The expressions for the optimum state estimate and for the covariance matrix of the error in the estimate are given by Eqs. 16 and 17 in which t is replaced by i.

The linear discrete time case was first solved (in a different form) by Magill (7). It is presented here as an application of a more general result, Eq. 12, and to provide a comparison to the continuous time case.

EXAMPLE 2

The problem of the detection of known or unknown signals can be treated as an application of the results of this paper. Define the set of likelihood functions $\{L_i(t), 1<i<q\}$ such that:

$$L_i(t) = \ln \left\{ p_i(o) E[e^{\Theta_i(t)} \mid O_t, \alpha = k_i] \right\}, \qquad [20]$$

$L_i(t)$ is seen to be the logarithm of numerator of the right hand side of Eq. 10.

Suppose $\{h[x(t),t,k_i], 1 \leq i \leq q\}$ represents a set of (q) signals, one of which has been transmitted. The received waveform is $y(s), s \in [t_o, t]$. The signal detection problem is to determine which signal was transmitted upon receipt of y(s). Clearly, the signal corresponding to the maximum $L_i(t)$ is the most likely to have been transmitted. A device which computed $L_i(t)$ and then selected the signal corresponding to the maximum $L_i(t)$ would be an "optimum" detector.

Application of Ito's theorem to Eq. 20 gives the set of stochastic differential equations:

$$dL_i(t) = \hat{h}_i' R^{-1}(t) dy(t) - \frac{1}{2} \hat{h}_i' R^{-1}(t) \hat{h}_i, \qquad [21]$$

which describe the likelihood function.

In the case of known signals,

$$h[x(t),t,k_i] \equiv h_i(t), \qquad [22]$$

Eq. 21 can be rewritten in integral form:

$$L_i(t) = c_i(t) + \int_{t_o}^{t} h_i'(s) R^{-1}(s)\, dy(s) \qquad [23]$$

where;

$$c_i(t) = \ln \left\{ p_i(o) \right\} - \frac{1}{2} \int_{t_o}^{t} h_i'(s) R^{-1}(s) h_i(s)\, ds. \qquad [24]$$

In the scalar, time invariant case, $[R(s) = R]$, Eqs. 23 and 24 correspond to the well known matched filter result.

For random signals, signal estimation must usually be done in order to determine the most likely signal. The signal estimates, $\hat{h}_i$, are used in Eq. 21, or an equivalent integral equation, to determine $L_i(t)$.

CONCLUSIONS

Exact equations which describe the conditional density of the state of a system described by nonlinear stochastic differential equations having an unknown discretely distributed parameter have been derived. The density is conditioned on a continuous vector measurement which consists of the sum of a nonlinear functional of the state and a Wiener process. Parallel results for the discrete time case are also presented.

Two examples demonstrating the application of these results to the estimation and detection problems are presented. Further illustrations in the detection area could have been given, but due to lack of space, they were not. For example, the problem of detection of signals in colored noise can be handled by "whitening" the measurement

noise. In short, many of the known results of detection theory can be easily derived using the equations obtained in this paper.

ACKNOWLEDGEMENT

The author would like to thank Prof. R. A. Hoddad of the Polytechnic Institute of Brooklyn for many helpful suggestions, and Mr. I. Citron of Singer-General Precision, Inc. who initially suggested a related problem.

REFERENCES

1. Stratonovich, R.I., "Conditional Markov Processes", Theory Prob. Appl 5 (1960), 156-178.

2. Kushner, H.J., "On the Dynamical Equations of Conditional Probability Density Functions With Applications to Optimal Stochastic Control Theory", J. Math Anal. Appl 8 (1964), 332-334.

3. Kushner, H.J., On the differential equation satisfied by conditional densities of Markov processes, J. SIAM Control Ser. A (1964), 106-119.

4. Kushner, H.J., "Dynamical Equations for Optimum Nonlinear Filtering", J Differential Equations, Vol. 3, pp 79-190, 1967.

5. Bucy, R.S.,"Nonlinear Filtering", IEEE Trans. Automatic Control, Vol. AC-10, p. 198, 1965.

6. Bucy, R.S. and Joseph, P.D., "Filtering for Stochastic Processes with Applications to Guidance", Interscience, New York, 1968.

7. Magill, D.T., "Optimal Adaptive Estimation of Sampled Stochastic Processes", IEEE Trans. Automatic Control, Vol. AC-10, pp 434-439, 1965.

8. Ito, K., "On Stochastic Processes", Lecture Notes, Tata Institute for Fundamental Research, Bombay, 1961.

9. Wonham, W.M., "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering", J. SIAM Control Ser. A (1964), pp 106-109.

10. Doob, J.L., "Stochastic Processes", Wiley, New York, 1953.

APPENDIX A PROOF OF THEOREM 3

Let $t > o$ be fixed; for each positive integer (n), define the partition of $[t_o, t]$:

$$t_o = t_o^n \leq t_1^n \ldots . t_n^n = t;$$

$$\delta y_i^n = \int_{t_{i-1}}^{t_i} dy(s);$$

$$\delta H_i^n(k_j) = \int_{t_{i-1}}^{t_i} h[x(s), s, k_j]\, ds;$$

$$\delta \hat{w}_i^n = \int_{t_{i-1}}^{t_i} R^{1/2}(s)\, dw(s).$$

Then:

$$\delta y_i^n = \delta H_i^n(k_j) + \delta \hat{w}_i^n .$$

Let O_n denote the minimum σ - field determined by δy_i^n $i=1,..n$. The δw_i^n are independent normal random variables with covariance:

$$S_i^n = \int_{t_{i-1}}^{t_i} R(s)\, ds.$$

Let $\mathcal{H}_n$ denote the minimum σ - field determined by $\delta H_i^n(k_j)$ $i=1,..n$ then:

$$p[\delta y_i^n, i=1,..n \mid \mathcal{H}_n, \alpha=k_j] = C \exp -\tfrac{1}{2} \sum_{i=1}^{n} \| \delta y_i^n - \delta H_i^n(k_j) \|^2 (S_i^n)^{-1},$$

where C is a normalizing constant.

Define:

$$p_j^n(t) = p\{ \alpha = k_j \mid O_n \} .$$

An expression for this conditional probability is found by applying Bayes' rule:

$$p_j^n(t) = \frac{p_j(o)\, p[\delta y_i^n, i=1,..n \mid \alpha = k_j]}{\sum_{k=1}^{q} p_k(o)\, p[\delta y_i^n, i=1,..n \mid \alpha = k_k]} . \qquad [A-1]$$

The conditional density $p[\delta y_i^n, i=1..n \mid \alpha = k_j]$ can be written as $p[\delta y_i^n, i=1..n \mid \alpha=k_j] = \int_{-\infty}^{\infty} p[\delta y_i^n, i=1..n \mid \alpha=k_j, \mathcal{H}_n]\, p[\delta H^n(k_j)]\, d\,\delta H^n$

where:

$$\delta H^n = \{ \delta H_i^n, i \leq n \}.$$

Let $\bar{x}(t)$ be independent of $x(t)$ but have the same distribution. Define:

$$\delta \bar{H}_i^n(k_j) = \int_{t_{i-1}}^{t_i} h[\bar{x}(s), s, k_j]\, ds.$$

Then $p[\delta y_i^n, i=1..n \mid \alpha=k_j]$ can be written as $p[\delta y_i^n, i=1..n \mid \alpha=k_j] = CE\{ \exp -\tfrac{1}{2} \sum_{i=1}^{n} \| \delta y_i^n - \delta \bar{H}_i^n(k_j) \|^2 (S_i^n)^{-1} \mid O_n, \alpha = k_j \}$. [A-2]

Substituting Eq. A-2 into Eq. A-1 and multiplying the numerator and denominator by $\exp \frac{1}{2} \sum_{i=1}^{n} \| \delta y_i^n \|^2_{(S_i^n)^{-1}}$, gives:

$$p_j^n(t) = \frac{p_j(o)\, E\left\{ e^{\Theta_j^n(t)} \mid O_n ,\ \alpha = k_j \right\}}{\sum_{k=1}^{q} p_k(o)\, E\left\{ e^{\Theta_j^n(t)} \mid O_n , \alpha = k_j \right\}} , \qquad [A\text{-}3]$$

where:

$$\Theta_j^n(t) = -\frac{1}{2} \sum_{i=1}^{n} \| \delta \bar{H}_i^n(k_j) \|^2_{(S_i^n)^{-1}} + \sum_{i=1}^{n} [\delta \bar{H}_i^n(k_j)]' (S_i^n)^{-1} \delta y_i^n .$$

Convergence of $p_j^n(t)$ to $p_j(t)$ must be proved. A similar proof in Wonham [9] is followed here. By definition of the stochastic integral [10, Chapter IX] and the continuity of $R^{-1}(s)$ and $x(s)$, $s \in [t_o, t]$, it follows that:

$$\underset{n\to\infty}{\text{l.i.m.}} \sum_{i=1}^{n} \delta \bar{H}_i^n(k_j) (S_i^n)^{-1} \delta y_i^n = \int_{t_o}^{t} h[\bar{x}(s), s, k_j]' R^{-1}(s)\, dy(s)$$

$$\lim_{n\to\infty} \sum_{i=1}^{n} \| \delta \bar{H}_i^n(k_j) \|^2_{(S_i^n)^{-1}} = \int_{t_o}^{t} \| h[\bar{x}(s), s, k_j] \|^2_{R^{-1}(s)} ds ,$$

and hence $\Theta_j^n(t) \to \Theta_j(t)$ in the mean. It remains only to prove that:

$$p_j(t) = \underset{n\to\infty}{\text{l.i.m.}}\ p_j^n(t) .$$

Let $\{\tau_\nu, \nu = 1 \ldots N\}$ be a subset of $\{t_i, i = 1 \ldots n\}$. Let Ω be the event $\{y(\tau_\nu) \in A_\nu, \nu = 1 \ldots N\}$, then:

$$\int_{\Omega} p_j^n(t)\, dP = P\{\Omega, \alpha = k_j\}, \qquad [A\text{-}4]$$

which holds for $n \to \infty$.

Since Eq. A-4 holds for every Ω of the form described, it has been shown that $\underset{n\to\infty}{\text{l.i.m.}}\ p_j^n(t)$ has the defining properties of the conditional probability $p_j(t)$ which completes the proof.

In order to prove Corollary 3, note that the joint process $x(t)$, Θ_j, $j = 1 \ldots q$ is Markov. Eq. 10 is given by

$$p_j(t) = \frac{p_j(o)\, E\left\{ e^{\Theta_j(t)} \mid O_t ,\ \alpha = k_j \right\}}{\sum_{k=1}^{q} p_k(o) E\left\{ e^{\Theta_j(t)} \mid O_t , \alpha = k_j \right\}} = \frac{N_j}{D} .$$

It follows from Ito's theorem that:

$$dN_j = p_j(o)\, E\left\{ h'[\bar{x}(t), t, k_j]\, e^{\Theta_j(t)} \mid O_t , \alpha = k_j \right\} R^{-1}(t)\, dy(t) . \qquad [A\text{-}5]$$

Note that:

$$dD = \sum_{j=1}^{q} dN_j . \qquad [A\text{-}6]$$

Apply Corollary 1 to Eqs. A-5 and A-6 which gives:

$$dN_j = N_j \hat{h}_j' R^{-1}(t)\, dy(t) , \qquad [A\text{-}7]$$

$$dD = D \hat{h}' R^{-1}(t)\, dy(t) . \qquad [A\text{-}8]$$

where $\hat{h} = \sum_{j=1}^{q} p_j(t) \hat{h}_j(t)$. Eq. 11 then follows from Corollary 2.

A GENERAL APPROACH TO DISCRETE SYSTEM IDENTIFICATION

K.L. SURYANARAYANAN, A.C. SOUDACK and S.G. RAO
Electrical Engineering Department
University of British Columbia, Vancouver 168, B.C., Canada

ABSTRACT

A systematic on-line noniterative procedure of identification of discrete-time systems is presented.

A noise free system, whose difference equation (d.e.) is given below, is considered for identification:

$$K^T\bar{\phi}(j) = \sum_{i=1}^{N} k_i \phi_i(r(n),y(n)) = \sum_{i=1}^{N} k_i \phi_i(j) = 0;\ n \leq j;\ j=1,2..... \qquad (1)$$

where k_i are unknown coefficients, ϕ_i are known functions of measurable quantities, r(n) (the system input) and y(n) (the system output). A model is chosen with the d.e.

$$\hat{K}^T\hat{\bar{\phi}}(j) = \sum_{i=1}^{N} \hat{k}_i \phi_i(r(n),\hat{y}(n)) = \sum_{i=1}^{N} \hat{k}_i \hat{\phi}_i(j) = 0 \qquad (2)$$

where cap $\hat{}$ refers to model quantities. It is required to adjust $\hat{K}$ such that $\hat{K} \to K$.

The general procedure consists of the following steps:
(i) Choose a positive definite function V of a suitable error variable.
(ii) Choose the increment in $\hat{K}$, $\Delta\hat{K}=hF$ (h, real, and F, vector) and minimize ΔV by proper choice of h.
(iii) Complete the design by choosing F such that $\Delta\hat{K}$ is a vector of known factors and check for convergence of $\hat{K}$ to K.

Identification Error Method

(i) Scalar function $V_I(j) = E^T(j)ME(j)$; M, symmetric, positive definite.

$$E(j) \triangleq K - \hat{K}(j);\ \text{and}\ \Delta V_I(j) = V_I(j+1) - V_I(j) \qquad (3)$$

(ii) Following step (ii), minimize ΔV_I to obtain

$$h = E^TMF/F^TMF;\ \text{and}\ \min \Delta V_I = -(E^TMF)^2/F^TMF \qquad (4)$$

dropping j for convenience.
(iii) Choosing $F = M^{-1}\bar{\phi}$ and using (1), (3), (4),

$$\Delta\hat{K} = E^T\bar{\phi}M^{-1}\bar{\phi}/(\bar{\phi}^TM^{-1}\bar{\phi}) = -\hat{K}^T\bar{\phi}M^{-1}\bar{\phi}/(\bar{\phi}^TM^{-1}\bar{\phi}) \qquad (5)$$

The convergence of $\hat{K}$ to K is found by examination of eqn. (5). Generally K is time varying so that $\Delta E(j) = \Delta K(j) - \Delta\hat{K}(j)$. Then eqn. (5) becomes,

$$\Delta E = -E^T\bar{\phi}M^{-1}\bar{\phi}/(\bar{\phi}^TM^{-1}\bar{\phi}) + \Delta K \qquad (6)$$

Eqn. (9) is analyzed using the concept of eventual asymptotic stability (e.a.s.). Liapunov stability does not apply here since E=0 is not an equilibrium point and d.e. (6) is nonautonomous. Note that e.a.s. implies $\hat{K}(j) \to K(j)$ as $j \to \infty$.

Theorem[1]: For the system and model of eqns. (1) and (2), if $\|K(j)\|$ is bounded and $\Delta K(j)=0$ for all j > some L, with at least one element of K known, and if the rank of the matrix,

$$\Psi(j,J_j) = [\bar{\phi}^p(j)\ \bar{\phi}^p(j+1)\ ....\ \bar{\phi}^p(j+J_j-1)] \qquad (7)$$

(where $\bar{\phi}^p(j) = [\phi_1\ \phi_2\ ....\ \phi_p]^T$; ϕ_1, ..., ϕ_p, respective multipliers of the unknown coefficients $k_1, k_2, ...k_p$) is p for all j and some J_j, then the origin of the system of eqn (6) with state E is e.a.s.

Output Error Method: The system and model d.e. are assumed to be given by

$$y(j) = \sum_{i=1}^{p} k_i \phi_i(j) = K^T \bar{\phi}; \text{ and } \hat{y}(j) = \sum_{i=1}^{p} \hat{k}_i \tilde{\phi}_i = \hat{K}^T \hat{\bar{\phi}} \qquad (8)$$

The output error, $e(j) \triangleq y(j) - \hat{y}(j)$. For the scalar function in this case, the choice of $V_0(j) = e^2(j)$ is made. An algorithm for $\Delta\hat{K}$ may be developed following the steps given earlier, giving:

$$\Delta\hat{K} = [K^T \bar{\phi}(j+1) - \hat{K}^T(j)\hat{\bar{\phi}}(j+1)]F/F^T\hat{\bar{\phi}}(j+1) \qquad (9)$$

If now the variation in K is also considered, $\Delta e(j) \triangleq e(j+1) - e(j)$, is given by

$$\Delta e(j) = -e(j) + \Delta K^T(j)\bar{\phi}(j+1) \qquad (10)$$

Since Liapunov stability is once again inapplicable to the above d.e., convergence is established by means of eventual stability.

Theorem[1]. The origin of d.e. (10) is e.a.s. provided $\|\Delta K(j)\|$ is bounded for all j and $\Delta K(j) \to 0$ for $j \to \infty$.

Note that e.a.s. implies that the model "tracks" the system output.

In implementing eqn (9), one assumes that the future quantities, $\hat{\bar{\phi}}(j+1)$, $\bar{\phi}(j+1)$ are known and $K^T\bar{\phi}(j+1)$ is available. These requirements are satisfied approximately by means of power series expansion at instant j. One possible choice for the remaining unknown F in eqn (9) is $F = M\bar{\phi}(j)$. Thus $\Delta\hat{K}$ is measurable.

Under suitable assumptions, output tracking implies that $\|E\| \to 0$[1].

Equation Error Method: For the system and model as given by eqn (8), the scalar function for minimization is chosen to be $V_E(j) = \varepsilon^2(j)$ [where the equation error $\varepsilon(j) \triangleq y(j) - \hat{K}^T(j)\bar{\phi}(j)$]. In this case, the algorithm for $\Delta\hat{K}$ is given by

$$\Delta\hat{K} = [K^T\bar{\phi}(j+1) - \hat{K}(j)\bar{\phi}(j+1)]F/F^T\bar{\phi}(j+1) \qquad (11)$$

Let $F = \bar{\phi}(j+1)$. With this choice, $\Delta\hat{K}$ is measurable since $\bar{\phi}(j+1)$ and $K^T\bar{\phi}(j+1)$ are known approximately. If now K is considered time varying, then eqn (11) becomes,

$$\Delta E(j) = -E^T(j)\bar{\phi}(j+1)F/(F^T\bar{\phi}(j+1)) + \Delta K(j) \qquad (12)$$

The convergence of E to origin is established by means of a theorem.

Theorem[1]: The system of eqn (12) is e.a.s. if $\|K(j)\|$ is bounded and $\Delta K(j) = 0$ for all $j > L$ and the rank of $\Psi(j+1, J_j)$ is p for all j and some J_j.

A detailed discussion of the above methods will be found in Suryanarayanan[1].

ACKNOWLEDGEMENT

This research was supported by NRC Grant No. 67-3138. One of the authors (K.L.S.) gratefully acknowledges the scholarship awarded to him by the National Research Council of Canada during 1966-1969.

REFERENCES

1) K.L. Suryanarayanan, "Identification of Nonlinear Sampled-data Systems", Ph.D. Thesis, Dept. of Elec. Eng., Univ. of British Columbia, Canada. (in preparation)

A NON-PARAMETRIC RECURSIVE IDENTIFICATION ALGORITHM FOR FINITE-MEMORY LINEAR SYSTEMS*

Y. BAR-SHALOM** and S.C. SCHWARTZ**
Princeton University, Princeton, New Jersey 08540

ABSTRACT

The problem of identifying a linear, finite-memory, discrete-time system with noisy input and output is considered. A recursive algorithm is proposed which gives asymptotically unbiased estimates of the system parameters. The procedure does not require any specific knowledge of the noise statistics, in contrast to other proposed algorithms which generally need some knowledge of the statistics in order to obtain unbiased parameter estimates.

Introduction

The problem of identification of linear systems in noisy environments has received a great deal of attention [1-5]. The simple computational requirements of recursive estimation techniques makes them attractive for on-line identification purposes. Recently, stochastic approximation algorithms have been proposed for the identification of linear systems in a noisy environment. Sakrison [6] considered the identification of a system modelled by a Laplace transform transfer function. The algorithm proposed by Saridis and Stein [7] is for discrete-time systems represented in state-space or by z-transform transfer function, while Holmes [8] characterized the unknown linear system by a pulse response.

A common feature of these techniques is the requirement that the correlation function (in [6]) or the variances (in [7,8]) of the noises be known. In [6] the correlation functions of input and output noises are needed to form the gradient of the error function used in the algorithm. The variances of both input and output noises in [7] and the variance of the input noise in [6] are required so as to add a correction term to the recursive procedure so that the estimates of the system

* Research supported by National Science Foundation Grant NSF-GK-1532 and Air Force Office of Scientific Research under AFOSR Grant 1333-67.
** Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08540

parameters are asymptotically unbiased. In another paper [13] Saridis and Stein present an algorithm which only requires knowledge of the variance of the input measurement noise.

In this paper, an algorithm is presented which does not require knowledge of the noise variances and for which the estimates converge to the true system parameters both in mean-square and with probability one.[1] Another departure from previous techniques is that this procedure uses a deterministic orthogonal sequence of inputs, as contrasted to the usual stationary random process as input. The algorithm is a stochastic approximation technique with certain characteristics of a well-known error-correcting procedure used in pattern recognition problems (Nilsson [9], p.90).

The Algorithm

A scalar input, scalar output linear system with finite memory is characterized by an N-vector

$$\underline{h}' = \left[h_0, h_1, \ldots, h_{N-1}\right] \tag{1}$$

where prime denotes transposed.

Denoting by x_k the input at time k, the output at time j is

$$y_j = \sum_{i=0}^{N-1} h_i x_{j-i} = \underline{h}' \, \underline{x}_j \tag{2}$$

where

$$\underline{x}_j' = \left[x_j, x_{j-1}, \ldots, x_{j-N+1}\right] \tag{3}$$

is the "sliding vector" of the last N inputs.

If input and output noises are present, the system is described by the equation

$$y_j = \underline{h}' \, (\underline{x}_j + \underline{n}_j) + m_j \tag{4}$$

with the input noise vector

$$\underline{n}_j' = \left[n_j, n_{j-1}, \ldots, n_{j-N+1}\right] \tag{5}$$

and output noise m_j. It will be assumed that both noises are

[1] Two other nonparametric algorithms having these convergence properties have been proposed by Saridis and Stein [7]. However, their technique requires the inversion of a matrix which may be singular.

zero mean, uncorrelated, and have bounded variances

$$E\left[n_j n_k\right] \leq \sigma_n^2 \delta_{jk},\ E\left[m_j m_k\right] \leq \sigma_m^2 \delta_{jk};\ j,k=1,2,\ldots \tag{6}$$

$$E\left[n_j m_k\right] = 0;\ j,k=1,2,\ldots \tag{7}$$

The problem is to find an estimate (at time j)

$$\underline{\hat{h}}_j' = \left[\hat{h}_{0,j},\ \hat{h}_{1,j},\ \ldots,\ \hat{h}_{N-1,j}\right] \tag{8}$$

which will converge to h, i.e.

$$\hat{h}_{i,j} \to \hat{h}_i = h_i \text{ as } j\to\infty;\ i=0,1,\ \ldots,\ N-1. \tag{9}$$

It is assumed that the memory of the system is not larger than N.

The algorithm used is the following:

$$\underline{\hat{h}}_{j+1} = \underline{\hat{h}}_j + a_j \left[y_j - \underline{\hat{h}}_j' \underline{x}_j\right] \frac{\underline{x}_j}{\|\underline{x}_j\|^2} \tag{10}$$

where $\|\cdot\|$ is the Euclidean norm.

Nagumo and Noda [10] used the above algorithm with a_j constant and proved convergence for the case where both the input and output signals can be measured exactly.

The other stochastic approximation schemes differ from this one in that they do not "normalize" the input. As will be seen below this scaling will play an important role. In addition, as mentioned in the introduction, all the previous schemes used a stationary random process as input while here the input will be chosen as a deterministic sequence (see (18) and (19)).

Using (1), the (j+1)-th estimate of the pulse response is

$$\underline{\hat{h}}_{j+1} = \underline{\hat{h}}_j + a_j\left[\underline{h}'\underline{x}_j + \underline{h}'\underline{n}_j + m_j - \underline{\hat{h}}_j'\underline{x}_j\right] \frac{\underline{x}_j}{\|\underline{x}_j\|^2} \tag{11}$$

Define the error:

$$\underline{e}_{j+1} = \underline{\hat{h}}_{j+1} - \underline{h} \tag{12}$$

The dynamical equation which governs the error vector is:

$$\underline{e}_{j+1} = \Phi_j \underline{e}_j + \Gamma_j \underline{n}_j + \underline{\delta}_j m_j \tag{13}$$

where

$$\Phi_j = I - a_j \frac{\underline{x}_j \underline{x}_j'}{||\underline{x}_j||^2} \quad ; \; I = \text{identity matrix} \tag{14}$$

$$\Gamma_j = a_j \frac{\underline{x}_j \underline{h}'}{||\underline{x}_j||^2} \tag{15}$$

$$\delta_j = a_j \frac{\underline{x}_j}{||\underline{x}_j||^2} \tag{16}$$

The Choice of the Input Sequence

From (13) one can see that a necessary condition for the error to converge to zero is that Φ_j be a contraction mapping. Since one eigenvalue of Φ_j is $\lambda_0 = 1 - a_j$, corresponding to the eigenvector $u_{0,j} = \underline{x}_j$, in order to have a contraction, the following condition must be satisfied (see also [9]).

$$0 < a_j < 2 \tag{17}$$

All other eigenvalues (corresponding to the remaining eigenvectors, orthogonal to $\underline{x}_j$) are equal to 1. Hence the "contraction" (i.e. decrease in the error) at time j will occur only along the direction of $\underline{x}_j$. Therefore, the sequence $\langle x_j \rangle$ is to be chosen such that $\underline{x}_j$ will repeatedly span the whole N-dimensional Euclidean space. Observe that the norm $||\underline{x}_j||$ does not affect the size of contraction but only the disturbance due to the noises (see (13)). The larger $||\underline{x}_j||$ is the smaller is this disturbance. Since in physical systems the amplitude of the input must be limited, the input sequence will be chosen so that $\underline{x}_j$ will have a fixed, maximum norm. Consequently, the input sequence is chosen such that $\{\underline{x}_j\}_{j=1}^{N}$ forms a set of orthogonal vectors which uniformly spans the N-dimensional space which, by assumption, contains the unknown system vector $\underline{h}$. One sequence having these properties is:

$$x_j = \begin{cases} U>0 & j=kN+1 \quad k=0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases} \qquad j=1,2,\ldots \tag{18}$$

i.e.

$$x_j = U\delta_{j,kN+1} \qquad k=0,1,2,\ldots \quad ; \; j=1,2,\ldots \tag{19}$$

This will guarantee that all the components of the error vector converge to zero.

The sequence $\langle a_j \rangle$ appearing in the algorithm (11) will be chosen as follows:

$$a_j = \alpha_k \qquad kN+1 \leq j < (k+1)N \qquad k=0,1,2,\ldots \tag{20}$$

i.e. for N consecutive sampling times a_j is constant.

Since x_j is known there is no need to measure it. In [13], the input measurement noise introduces a bias, which can only be eliminated by adding a correction term to thealgorithm. For the present algorithm, inaccuracies in the application of the input are reflected in the input noise.

Proof of Convergence

Defining the index $j=kN+i$; $i=1,2,\ldots,N$; $k=0,1,2,\ldots$ equation (13) becomes

$$\underline{e}_{kN+i+1} = \Phi_{kN+i}\underline{e}_{kN+i} + \Gamma_{kN+i}\underline{n}_{kN+i} + \underline{\delta}_{kN+i}m_{kN+i} \tag{21}$$

where

$$\Phi_{kN+i} = I - \alpha_k \frac{\underline{x}_i\underline{x}_i'}{||\underline{x}_i||^2} \tag{22}$$

$$\Gamma_{kN+i} = \alpha_k \frac{\underline{x}_i h'}{||\underline{x}_i||^2} \tag{23}$$

$$\delta_{kN+i} = \alpha_k \frac{\underline{x}_i}{||\underline{x}_i||^2} \tag{24}$$

Assume, for simplicity, U=1, and recall that the inputs were chose so that

$$\underline{x}_\mu' \underline{x}_\nu = \delta_{\mu\nu} , \qquad \mu,\nu = 1,2,\ldots,N \tag{25}$$

Then the projection of the error vector on $\underline{x}_\mu$ is given by

$$\underline{x}_\mu' \underline{e}_{kN+i+1} = \underline{x}_\mu' \underline{e}_{kN+i} \qquad \mu \neq i \tag{26}$$

and for $\mu=i$,

$$\underline{x}_\mu' \underline{e}_{kN+\mu+1} = (1-\alpha_k)\underline{x}_\mu' \underline{e}_{kN+\mu} + \alpha_k \underline{n}_{kN+\mu}' \underline{h} + \alpha_k m_{kN+\mu} \tag{27}$$

The mean square value of these projections is:

$$E\left[(\underline{x}_\mu' \underline{e}_{kN+i+1})^2 \right] = E\left[(\underline{x}_\mu' \underline{e}_{kN+i})^2 \right] \qquad \mu \neq i \tag{28}$$

and for $\mu=i$

$$E\left[(\underline{x}_\mu' \underline{e}_{kN+\mu+1})^2\right] = (1-\alpha_k)^2 E\left[(\underline{x}_\mu' \underline{e}_{kN+\mu})^2\right]$$

$$+ \alpha_k^2 E\left[\underline{n}_{kN+\mu}' \underline{h}\, \underline{h}'\, \underline{n}_{kN+\mu}\right]$$

$$+ \alpha_k^2 E\left[m_{kN+\mu}^2\right] + 2\alpha_k(1-\alpha_k) E\left[\underline{x}_\mu' \underline{e}_{kN+\mu} \underline{h}' \underline{n}_{kN+\mu}\right] \qquad (29)$$

From the assumptions on the noises, the other two crossterms in (29) are zero and not shown. Since $\underline{e}_{kN+\mu}$ and $\underline{n}_{kN+\mu}$ are dependent, it is not obvious that the remaining crossterm in equation (29) is zero. However, from (26)

$$\underline{x}_\mu' \underline{e}_{kN+\mu} = \underline{x}_\mu' \underline{e}_{kN+\mu-1} = \cdots = \underline{x}_\mu' \underline{e}_{(k-1)N+\mu+1} \qquad (30)$$

and since $\underline{e}_{(k-1)N+\mu+1}$ is independent of $\underline{n}_{kN+\mu}$, it follows from (6) and (21) that this crossterm will also be zero. Therefore, from equations (28) and (29), one can write for the μ-th component of the error

$$E\left[(\underline{x}_\mu' \underline{e}_{(k+1)N})^2\right] = (1-\alpha_k)^2 E\left[(\underline{x}_\mu' \underline{e}_{kN})^2\right]$$

$$+ \alpha_k^2 E\left[\underline{n}_{kN+\mu}' \underline{h}\, \underline{h}'\, \underline{n}_{kN+\mu}\right] + \alpha_k^2 E\left[m_{kN+\mu}^2\right] \qquad (31)$$

Denote by λ the greatest eigenvalue of the $N \times N$ matrix $\underline{h}\, \underline{h}'$. Using (6), the mean-square values of the projections of the error are bounded by:

$$E\left[(\underline{x}_\mu' \underline{e}_{(k+1)N})^2\right] \le (1-\alpha_k)^2 E\left[(\underline{x}_\mu' \underline{e}_{kN})^2\right] + \alpha_k^2 \sigma^2; \quad \mu = 1,2,..,N. \qquad (32)$$

where

$$\sigma^2 = N\lambda \sigma_n^2 + \sigma_m^2 \qquad (33)$$

It is well-known (see e.g. [11]) that under the following conditions

$$\sum_1^\infty \alpha_k = \infty \qquad (34) \qquad \sum_1^\infty \alpha_k^2 < \infty \qquad (35)$$

the left-hand side of (32) satisfies:

$$\lim_{k\to\infty} E\left[(\underline{x}_\mu' \underline{e}_{kN})^2\right] = 0 \qquad \mu = 1,2,\ldots,N \qquad (36)$$

The choice of a_j as in (20) yields the same rate of convergence for all the components of the error since each component decreases once during N sampling times and the correction

coefficient is constant for the same period. Since the set $\{\underline{x}_\mu\}_{\mu=1}^{N}$ spans the whole space, the error vector goes to zero i.m.s. Furthermore, as can be seen from (32), condition 5.1 of Dvoretzky [12] is satisfied, and therefore the error also converges to zero with probability one. Results of Dvoretzky also can be used to find the number of iterations needed to obtain the error below any given positive quantity.

Fig. 1 shows the normalized error in the parameter estimates for a system with a 10-term impulse response (N=10). The upper curve is for a strong noise; both input and output noises have a variance equal to the norm of the sliding input vector. The lower curve is for relatively weak noise. It is seen that in the first case, where, practically everything is "buried" in the noise one can identify $\underline{h}$ to within a few percent of error. The initial value for the iteration was $\hat{\underline{h}}_0=\underline{0}$

If the actual dimension of $\underline{h}$ is larger than the assumed value, the neglected terms will affect the results by introducing a bias. For example, if one assumes N=2 while actually N=3, the input sequence (19) results in the following output sequence:

input:	1	0	1	0	1	0	1 ...
output:	h_0	h_1	h_0+h_2	h_1	h_0+h_2	h_1	h_0+h_2 ...

h_0 has a bias equal to the neglected term, h_2. If the sign of the nonzero input alternates, h_0 will have a bias equal to $-h_2$. Now consider alternating the sign of the nonzero input every 2N sampling times, i.e.,

input:	1	0	1	0	-1	0	-1 ...
output:	h_0	h_1	h_0+h_2	h_1	$-h_0+h_2$	$-h_1$	$-h_0-h_2$...

Then, instead of measuring h_0, the values observed alternate between h_0+h_2 and h_0-h_2. Hence, for this example, there will be no asymptotic bias in the estimates due to the averaging property of the stochastic approximation algorithm.

The following table shows the improvement due to alternating the sign of the input sequence. A model with memory length in the range 2-10 is used to identify a two-pole system whose actual pulse response has an infinite memory. In order to isolate the effect of neglected terms, noiseless observations were taken.

System transfer function $Y(z)/X(z)$	model memory length N	percent bias in the estimate of h_0*		
		Const. sign input	Alternating sign input	
			each N	each 2N
$\frac{1}{(1-0.7z^{-1})(1-0.5z^{-1})}$	2	251.8	-65.2	-52.6
	3	146.3	-61.6	-32.4
	5	62.25	-42.8	- 8.71
	10	9.82	- 9.37	- 0.24

* 1000 observations

Conclusion

An algorithm has been presented which gives an asymptotically unbiased estimate of the pulse response of a discrete-time, finite-memory, linear system. The procedure possesses the characteristics of both stochastic approximation and a pattern recognition scheme. Unlike other methods, where the input is a sampled stationary random process, here it is chosen as a deterministic sequence. The algorithm is non-parametric since the only assumptions about the input and output noises are that they be white, uncorrelated and have finite variance. The error in the estimates converges to zero both in mean square and with probability one. A modified procedure has been discussed which reduces the bias introduced when the system memory is larger than that of the model.

References

1. R.E. Kalman, "Design of a Self-optimizing Control System", Trans. ASME, J. Basic Engrg., ser. D, Vol. 80, pp. 468-478, February 1958.

2. H.J. Kushner, "A Simple Iterative Procedure for the Identification of the Unknown Parameters of Linear Time Varying Discrete System", Trans. ASME, J. Basic Engrg., ser. D, Vol. 85, pp. 227-235, June 1963.

3. M.J. Levin, "Estimation of a System Pulse Transfer Function in the Presence of Noise", IEEE Trans. Automatic Control, Vol. AC-9, pp. 229-235, July 1964.

4. Y.C. Ho and R.C.K. Lee, "Identification of Linear Dynamic Systems," J. Information and Control, Vol. 8, pp. 93-110, 1965.

5. A.E. Rogers and K. Steiglitz, "Maximum Likelihood Estimation of Rational Transfer Function Parameters," IEEE Trans. Automatic Control, (short papers), Vol. AC-12, pp.594-597, October 1967.

6. D.J. Sakrison, "The Use of Stochastic Approximation to Solve the System Identification Problem", IEEE Trans. Automatic Control, Vol. AC-12, pp. 563-567, October 1967.

7. G.N. Saridis and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-time System Identification" IEEE Trans. Automatic Control, Vol. AC-13, pp. 515-523, October 1968.

8. J.K. Holmes, "System Identification from Noise-corrupted Measurements," J. of Optimization Theory and Applications, Vol. 2, No. 2, pp. 102-116, 1968.

9. N.J. Nilsson, "Learning Machines", New York: McGraw-Hill, 1965.

10. J. Nagumo and A. Noda, "A Learning Method for System Identification", IEEE Trans. Automatic Control, Vol. AC-12, pp. 282-287, June 1967.

11. D.J. Sakrison, "Stochastic Approximation, a Recursive Method for Solving Regression Problems", in Advances in Communication Theory, Vol. 2, Ch. 2, A.V. Balakrishnan, Ed. New York: Academic Press, 1966.

12. A. Dvoretzky, "On Stochastic Approximation", Proc. 3rd Berkeley Symp. on Mathematical Statistics and Probability (1954-55), Vol. 1, pp. 39-55, Univ. of California Press, 1956.

13. G.N. Saridis and G. Stein, "A New Algorithm for System Identification", IEEE Trans. Automatic Control, Vol. AC-13, pp. 592-594, October 1968.

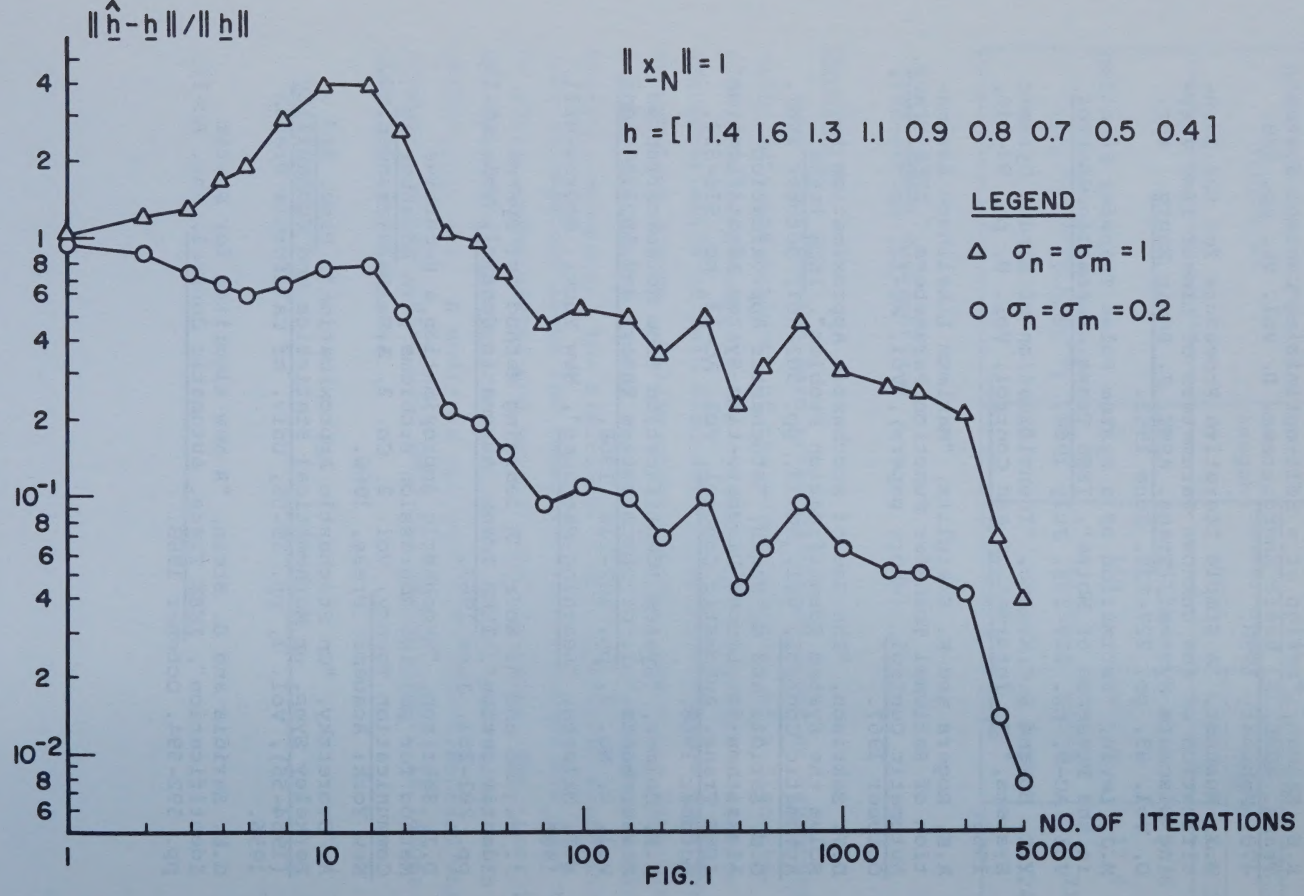

$\|\hat{h}-h\|/\|h\|$
$\|x_N\| = 1$
$h = [1\ 1.4\ 1.6\ 1.3\ 1.1\ 0.9\ 0.8\ 0.7\ 0.5\ 0.4]$
LEGEND
$\sigma_n = \sigma_m = 1$
$\sigma_n = \sigma_m = 0.2$
4
2
1
8
6
4
2
10^{-1}
8
6
4
2
10^{-2}
8
6
1
10
100
1000
5000
NO. OF ITERATIONS

FIG. 1

UNSUPERVISED LEARNING STRUCTURE AND PARAMETER ADAPTIVE PATTERN RECOGNITION. PART I: INDEPENDENT SIGNALLING, CONTINUOUS DATA

D. G. LAINIOTIS
Department of Electrical Engineering and Electronics Research Center
The University of Texas at Austin, Austin, Texas 78712

ABSTRACT

A Bayesian approach to optimal unsupervised learning, adaptive pattern recognition with continuous data and independent signalling is presented. Both structure and parameter adaptation are considered and specific optimal recursive adaptation algorithms are derived for gaussian process models and linear dynamics. Specifically, for the class of unsupervised learning, adaptive pattern recognition problems with hypothesis-conditional gaussian process models and linear dynamics, a "partition" theorem is derived that is applicable both to structure and parameter adaptation. The "partition" or "decomposition" theorem partitions the essentially nonlinear adaptive system into two parts, a linear and non-adaptive part consisting of recursive matched filters and a nonlinear part that incorporates the adaptive or learning nature of the optimal adaptive pattern recognition system. Extensions of the above result to the M-ary hypotheses case, where $M > 2$, as well as to the class of pattern recognition problems with both white and colored observation noise present, are given. In addition, a simple performance measure is intorduced for the on-line performance evaluation of the adaptive pattern recognition system. The on-line performance measure utilizes quantities readily available from the adaptive system and, hence, a minimum of additional computational effort is required for its evaluation.

SUMMARY

In this paper, Bayesian decision theory and state-variable linear dynamic modeling of gaussian random processes are applied to the problem of sequential, unsupervised learning, adaptive pattern classification to obtain the optimal-in the minimum-risk sense-solution in recursive form. The unsupervised learning problem or "learning without teacher" problem arises whenever knowledge of the correct classification of the sequence of previous observations is poor or nonexistent and the pattern recognition system is required to learn and adapt its structure to improve classification only on the basis of: a) A sequence of previously unclassified observations, i.e., while operating; and b) partial a-priori knowledge of the structure of the problem.

The assumed structural knowledge of the problem consists of a parametric model, where the relevant probability density functions have known functional forms which depend upon a set of unknown, constant parameters the number of which is also unknown. Specifically, gaussian signals and noise random processes are assumed, which are accurately described as the outputs of state-variable linear dynamic systems excited by white gaussian excitations. However, neither the signal and colored noise state-vector dimensionality, signal and noise system structure, respectively, nor the pertinent system parameter values are necessarily

known. A Bayesian approach to the problem is taken by assuming that the random processes involved are chosen at random by nature or an adversary from a collection of such processes with certain a-priori probabilities, not necessarily known either. They, the adaptive pattern recognition system, must learn both the order of the linear differential equations describing the time-evolution of the random processes involved, namely, the dimensionality of the signal and noise state-vectors (structure adaptation), as well as the coefficient of the differential equations (parameter adaptation).

Relevant Background

Early, recursive formulations of unsupervised learning pattern recognition are due to Fralick [1]. Specifically, he obtained the Bayes optimal solution for a sequence of binary decisions with sampled data where an unknown parameter (constant or time-varying Markov) pertains to one hypothesis and the similar multicategory problem where constant unknown parameters, pertaining to each of the hypotheses, are statistically independent; the parameters are restricted to take on finite sets of values. Patrick and Hancock [2] also obtained the Bayes optimal unsupervised solution for constant unknown parameters without requiring statistical independence between the class-characterizing parameters. In both Fralick [1] and Patrick and Hancock [2], the transmitter or pattern source was assumed to choose the categories in the sequence independently, and the observations were assumed to be conditionally independent. More recently, Hilborn and Lainiotis [3,4] extended and generalized previous results in many important aspects. Specifically, they have obtained the unsupervised learning multi-hypothesis, Bayes-optimal adaptive pattern recognition applicable to the parametric model with i) both constant and time-varying unknown parameters; ii) partially unknown probability laws of the hypotheses and time-varying unknown parameters sequences; iii) dependence of the observations on past, as well as present, hypotheses and parameters, i.e., for memory in observation medium or communication channel; and most significantly, iv) sequential dependencies in the observations arising from either (or both) dependence in the pattern or information source (context dependence) or in the observation medium (sequential measurement correlation) these dependencies being finite order Markov.

However, all of the above investigations were restricted to parameter adaptation only and to discrete data. In addition, due to previous formulation of the unsupervised learning problem and in particular due to the modeling used, it was not possible to utilize sufficient statistics in the calculation of the parameter-conditional likelihood ratios, as is done in supervised learning with consequent reduction in the computational and storage requirements. Moreover, no performance index has been proposed previously for the on-line performance evaluation of the adaptive recognition system.

New Results

In this paper, state-variable modeling of the random processes involved is utilized to obtain the following results [5,6]:

i) Introduction of the concept of structure adaptation in addition to parameter adaptation;

ii) derivation of optimal structure and parameter adaptive, unsupervised learning pattern recognition systems;

iii) for the class of pattern recognition problems with linear, gaussian models, a "partition" theorem, applicable to structure, as well as parameter adaptation is derived. The "partition" theorem partitions the nonlinear optimal adaptive pattern recognition system into a linear non-adaptive part consisting of recursive matched filters of the correlator-estimator form, and a nonlinear part that incorporates the adaptive nature of the system.

iv) Finally, a performance measure is introduced for the on-line evaluation of the adaptive recognition system performance. The measure, shown to be an estimate of an upper bound to the system probability of error, utilizes quantities readily available from the adaptive system and as such requires a minimum of additional computation for its evaluation.

REFERENCES

1. S. C. Fralcik, "Learning to Recognize Patterns without a Teacher," IEEE Transactions on Information Theory, Vol. IT-13, pp. 57-64, January, 1967.

2. E. A. Patrick and J. C. Hancock, "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Transactions on Information Theory, Vol. IT-12, pp. 362-372, July, 1966.

3. C. G. Hilborn and D. G. Lainiotis, "Optimal Unsupervised Learning Multi-category Dependent Hypotheses Pattern Recognition," IEEE Transactions on Information Theory, Vol. IT-14, pp. 468-470, May, 1968.

4. C. G. Hilborn and D. G. Lainiotis, "Unsupervised Learning Minimum Risk Pattern Classification for Dependent Hypotheses and Dependent Measurements," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-5, No. 2, pp. 109-115, April, 1969.

5. D. G. Lainiotis, "Unsupervised Learning Structure and Parameter Adaptive Pattern Recognition: Independent Signalling and Continuous Data," (Submitted for publication).

6. D. G. Lainiotis, "Supervised Learning Recursive Filters for Optimal Structure and Parameter Adaptive Pattern Recognition-Case II, Continuous Data," (Submitted for publication).

AN ADAPTIVE SCHEME FOR PATTERN RECOGNITION*

R. L. LUX AND D. C. LAI
University of Vermont
Burlington, Vermont

ABSTRACT

In many practical pattern recognition problems, only a small size of the initial learning samples is available. This paper describes an adaptive scheme which is designed to operate under such conditions. A quadratic discriminant function is used as the basic decision rule. Auxilliary criteria and a learning procedure are used to continually update the system parameters with each classification of new unlabeled samples. The system was successfully simulated and its performance was evaluated by using artificially generated samples as well as clinical laboratory data.

STATEMENT OF THE PROBLEM

A small number of labeled samples from each class of m classes is given. It is desired to design a practical pattern recognition scheme to classify any unlabeled member which is from one of these m classes. We have assumed that the members of the classes are normally distributed with unknown mean vector $\underline{M}_i$ and covariance matrix C_i for the i^{th} class. The a priori probabilities and the costs of classifications are assumed to be equal in the following solution.

THE RECOGNITION SCHEME AND EXPERIMENTAL RESULTS

The discriminant function derived by minimizing the average risk is

$$D_i(\underline{X}) = \ln|C_i| + (\underline{X} - \underline{M}_i)^t C_i^{-1} (\underline{X}-\underline{M}_i).$$

If $\underline{M}_i$ and C_i were known, then we would classify an unlabeled $\underline{X}$ as belonging to the class for which $D_i(\underline{X})$ is minimum. We derive initial boundaries between classes by using the unbiased estimates of $\underline{M}_i$ and C_i from the available labeled samples which are small in size. The initial decision regions are then partitioned into four zones as follows:

(A) Type A zones are enclosed by the 0.9 constant probability hyperellipsoids within which 90% of the class member distribution is expected to lie. The 0.9-contour described by the locus $\underline{X}_{0.9}$ for the i^{th} class is

$$\underline{X}_{0.9} = \underline{M}_i + R\sqrt{\lambda_j}\underline{E}_j$$

where $R^2 = (\underline{X}_{0.9} - \underline{M}_i)^t C_i^{-1} (\underline{X}_{0.9} - \underline{M}_i)$ and $\underline{E}_j$ is a normalized eigenvector of C_i and λ_j, the corresponding eigenvalue. (B) Type B zones are intersections of two or more A zones. (C) Type C zones consist of all the points outside A and B zones, but excluding strips along the boundaries. (D) Type D zones are described by points lying in strips along the boundaries, but not in A and B. The width of the strip is defined by a threshold which is determined empirically. These zones are shown in Figure 1 for a two-class, two-dimension case.

When an unlabeled sample $\underline{X}$ is given, we first make an initial decision based on the estimated boundaries and note the zone type in which the sample falls. If it falls in an A or C zone, the decision is assumed to be reliable. If a B zone is noted, the decision $X \in i^{th}$ class is made if a random variable y is observed and $y < b$, where y is uniformly distributed between 0 and 1, $b = \frac{n_j}{n_j+n_i}$, and $n_i = |D_i(\underline{X}) - \overline{D}_i|/\sqrt{v_i}$

*This research was supported in part by NASA Sustaining Grant NGR46-001-00852

with n_i similarly defined; $\bar{D}_i$ and V_i are, respectively, the mean and the variance of the random variable $D_i(\underline{X})$. If $\underline{X}$ falls in a D zone, no classification will be made. Whenever a decision $\underline{X} \in i^{th}$ class is made, the system parameters and the boundaries are updated according to the following formulas:

$$\hat{\underline{M}}_i' = \frac{1}{N+1}\,(N\hat{\underline{M}}_i + \underline{X})$$

$$\hat{C}_i' = \frac{1}{N}\,[(N-1)\hat{C}_i + (\underline{X} - \hat{\underline{M}}_i')\,(\underline{X} - \hat{\underline{M}}_i')^t]$$

$$\hat{\bar{D}}_i' = \frac{1}{N+1}\,[N\hat{\bar{D}}_i + D_i(\underline{X})]$$

$$\hat{V}_i' = \frac{1}{N}\,\{(N-1)\,\hat{V}_i + [\hat{\bar{D}}_i' - D_i(\underline{X})]^2\}$$

where $\hat{\underline{M}}_i$, $\hat{C}_i$, $\hat{\bar{D}}_i$, and $\hat{V}_i$ are the old parameters after classifying the N^{th} sample. The procedure is repeated when another unlabeled new sample is given.

This scheme is simulated and the performance is evaluated by using artificially generated samples and clinical laboratory data. The boundaries do converge to the theoretically calculated boundaries and average errors depending on the size of the initially labeled samples. It seems that there is a critical initial size for the process to converge. Some of the results are shown in Figures 2, 3, and 4.

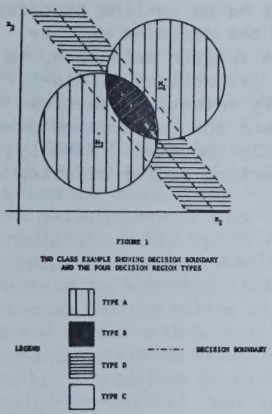


FIGURE 1

TWO CLASS EXAMPLE SHOWING DECISION BOUNDARY AND THE FOUR DECISION REGION TYPES

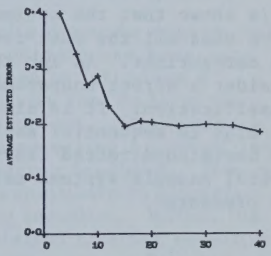
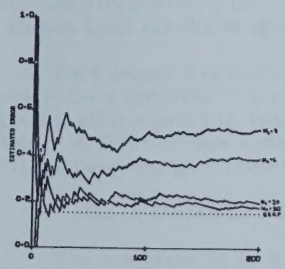


FIGURE 2

PERFORMANCE OF RECOGNITION SCHEME SHOWING ITS DEPENDENCE ON THE INITIAL NUMBER OF SAMPLES USED TO ESTIMATE THE CLASS PARAMETERS



PERFORMANCE CURVES SHOWING ESTIMATED ERROR DURING PROCESSING OF UNCLASSIFIED SAMPLES. THE INITIAL SAMPLE SIZE IS THE FAMILY PARAMETER

FIGURE 3

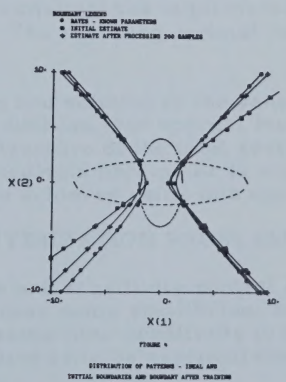


FIGURE 4

DISTRIBUTION OF PATTERNS - IDEAL AND INITIAL BOUNDARIES AND BOUNDARY AFTER TRAINING

STATISTICAL RECOGNIZER WITH ON-LINE FEATURE SELECTION AND ORDERING

Chi-hau Chen
Department of Electrical Engineering
Southeastern Massachusetts University
North Dartmouth, Mass. 02747

Summary*

A statistical recognizer consists of a receptor and a categorizer. For a finite number of pattern classes and available features (or feature subsets in vector measurements), the best performance is obtained by jointly optimizing the receptor and the categorizer. It is shown that this can be accomplished with an on-line feature selection followed by a minimum risk decision-making. The on-line scheme updates the selected features by using all available data - past and present. Pattern classification is based on the selected best features. In sequential pattern recognition, the first few feature observations are most effective in reducing the risk function. It is shown that the optimum feature ordering must be an on-line procedure which uses all the past feature observations plus the decision made by the categorizer. As for the nonparametric on-line feature selection, we consider a direct nonparametric estimation of the probability of misclassification. It is also proved that by properly selecting features, the bias in sequential estimation can be reduced and the convergence of the decision-directed learning system can be greatly improved. Finally, several example systems using on-line feature selection and/or ordering are presented.

---

* The full text will be available upon request.

DESIGN OF SENSITIVITY AND STATE VARIABLE FEEDBACK OPTIMAL CONTROL SYSTEMS

RONALD E. HIGGINBOTHAM
System Requirements and Simulation Section
Guidance and Control Systems Department
TRW Systems Group
Houston, Texas 77058

ABSTRACT

This paper considers the linear optimal control problem which has been reformulated so that sensitivity is controlled by the resulting control system configuration. It is shown that an optimal closed loop feedback control law can be synthesized if the problem is formulated by (1) augumenting the state vector by a set of sensitivity vectors, (2) augmenting the control vector by vectors which are functions of sensitivity variation, and (3) minimizing a quadratic cost functional in the augmented space. The resulting optimal feedback control law is a linear function of the state and the set of sensitivity vectors. A first-order example illustrates the performance obtainable using the design resulting from this approach.

I. INTRODUCTION

The problem of reducing state sensitivity of a dynamic system relative to plant parameter uncertainties in the time domain has been considered by several authors. The approach has been to introduce a constraint on state sensitivity by including sensitivity functions in the quadratic cost functional of the optimal control problem. The feedback structure obtained by this approach requires implementing a control law which is composed of both state and sensitivity vector feedback. However, to generate sensitivity vectors, the variation of sensitivity to uncertain plant parameters is explicitly needed as a forcing function. Within the formulation of the optimization problem, the variation of state sensitivity is computationally unrealizable because future values of sensitivity variation must be known (Reference 1). This problem of unrealizability can be circumvented by taking the variation of state sensitivity to be a sensitivity control signal which may be adjoined to the state-control signal. The augmented control vector is then included in the performance cost functional. In addition to an optimal state-control law, a set of optimal sensitivity-control laws is determined. Simultaneous implementation of both laws minimizes the cost functional. The resulting optimal control is a closed loop feedback law.

This paper presents the formulation and solution of the sensitivity constrained optimal control problem. In addition, the optimal feedback gain equations and the resulting control structure of the total system are given. A first-order example with one uncertain parameter is analyzed to illustrate the performance which can be achieved using this approach.

II. FORMULATION OF THE OPTIMIZATION PROBLEM

The problem is to find optimal state and sensitivity control laws which will keep the state of a given plant near some equilibrium state during real-time operation, while at the same time sensitivity to uncertain plant parameters is reduced. This objective is to be accomplished without excessive control-energy expenditure.

The plants to be controlled are deterministic, linear regulators. The mathematical model for a given regulator is assumed to contain some uncertain parameters whose n-dimensional state x(t) and r-dimensional control u(t) can be related by the following first-order, ordinary differential equations

$$\dot{x}(t) = A(t, \alpha)x(t) + B(t, \alpha)u(t), \qquad x(t_o) = x_o \tag{1}$$

where α is a constant vector of p plant parameters.

State sensitivity dynamics are established by first defining a set of sensitivity vectors (Reference 2)

$$v_j(t) = \frac{\partial x(t)}{\partial \alpha_j}, \quad (j=1, 2, \ldots, p) \tag{2}$$

Then from Equation (1), the sensitivity equations become:

$$\dot{v}_j(t) = A(t, \alpha_o)v_j(t) + \frac{\partial A(t, \alpha)}{\partial \alpha_j} x(t) + B(t, \alpha_o) \frac{\partial u}{\partial \alpha_j} + \frac{\partial B(t, \alpha)}{\partial \alpha_j} u(t) \tag{3}$$

where $v_j(t_o) = 0$ and α_o is the vector α evaluated using nominal values. Since the optimal state control law u(t) will be a function of state and sensitivity, expansion of $\partial u/\partial \alpha_j$ yields

$$\frac{\partial u(t)}{\partial \alpha_j} = [J_x u]v_j(t) + \sum_{i=1}^{p} \left[J_{v_i} u\right] \frac{\partial v_i}{\partial \alpha_j} \tag{4}$$

where $J_x u$ and $J_{v_i} u$ are Jacobian matrices (Reference 3). Substituting Equation (4) into (3) yields

$$\dot{v}_j(t) = [A(t, \alpha_o) + B(t, \alpha_o)J_x u]v_j(t) + \frac{\partial A}{\partial \alpha_j} x(t) + \frac{\partial B}{\partial \alpha_j} u(t)$$

$$+ B(t, \alpha_o) \sum_{i=1}^{p} \left[J_{v_i} u\right] \frac{\partial v_i}{\partial \alpha_j} \tag{5}$$

The terms $\partial v_i/\partial \alpha_j$ which appear in the preceding equation have been neglected by some authors (References 4, 5). This assumption is not valid in general. The problem formulation, as presented in Reference 1, results in an unrealizable control structure. To circumvent this, the $\partial v_i/\partial \alpha_j$ terms will be treated as sensitivity control vectors and as a result will be determined as linear feedback laws that minimize a specified cost functional. Let

$$m_j(t) \equiv \sum_{i=1}^{p} \left[J_{v_i} u\right] \frac{\partial v_i}{\partial \alpha_j} \tag{6}$$

where the dimension of each m_j is r. Also let z(t) be defined as the (p + 1) n-dimensional composite vector of the state and sensitivity vectors; that is, $z(t) = [x'(t), v_1'(t), \ldots, v_p'(t)]'$ where the prime (') denotes the transpose.

Equations (1) and (5) can be written as

$$\dot{z}(t) = A_1(t, \alpha)z(t) + B_1(t, \alpha)u_1(t), \qquad z(t_o) = z_o \tag{7}$$

where

$$u_1(t) = [u'(t), m_1'(t), \ldots, m_p'(t)]'.$$

A quadratic cost functional is assumed to correspond to the performance requirements. Thus,

$$J = \frac{1}{2} z'(t_f) Fz(t_f) + \frac{1}{2} \int_{t_o}^{t_f} [z'(t)Q(t)z(t) + u_1'(t)R(t)u_1(t)]\, dt \tag{8}$$

where the terminal time t_f is fixed and specified, F and Q(t) are symmetric positive semidefinite $(p + 1)n \times (p + 1)n$ matrices, and R(t) is a symmetric positive definite $(p + 1)r \times (p + 1)r$ matrix.

In summary, the problem is to find the optimal state control law u(t) and the set of optimal sensitivity control laws $m_j(t)$ that minimize the quadratic cost functional (8).

III. SOLUTION TO THE OPTIMIZATION PROBLEM

For the linear system (7) and the cost functional (8) along with the assumption that u(t), $m_1(t)$, ..., $m_p(t)$ are not restricted in magnitude, the optimal control laws are given by

$$u(t) = -R_1^{-1}\left[B'\left(K_{11}x + \sum_{j=1}^{p} K_{1k}v_j\right) + \sum_{j=1}^{p} \frac{\partial B'}{\partial \alpha_j}\left(K_{k1}x + \sum_{m=1}^{p} K_{kn}v_m\right)\right] \tag{9}$$

and

$$m_\mu(t) = -R_\nu^{-1}B'\left[K_{\nu 1}x + \sum_{j=1}^{p} K_{\nu k}v_j\right] \tag{10}$$

where $k = j + 1$, $n = m + 1$, and $\nu = \mu + 1$. Each R_ν is an $r \times r$ submatrix along the diagonal of R(t). Each $K_{ij}(t)$ is an $n \times n$ submatrix of the matrix K(t) which is the solution of the differential equation:

$$\dot{K}(t) + K(t)A_1(t, \alpha_o) + A_1'(t, \alpha_0)K(t) - K(t)M(t, \alpha_o)K(t) + Q(t) = 0 \tag{11}$$

where

$$M(t, \alpha_o) = B_1(t, \alpha_o)\, R^{-1}(t)\, B_1'(t, \alpha_o) \tag{12}$$

Equation (11) satisfies the boundary condition $K(t_f) = F$.

Figure 1 shows the structure of the optimal sensitivity and state controlled regulating system.

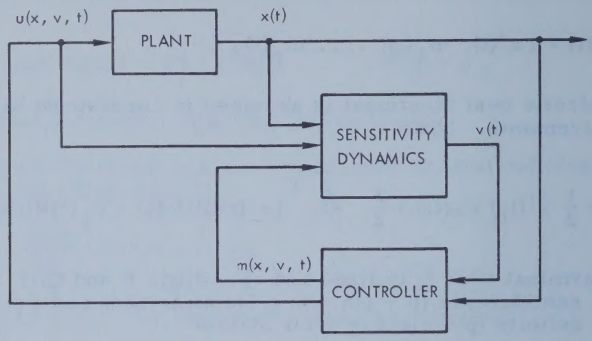


Figure 1. Structure of the Optimal Sensitivity and State Controlled Regulating System

IV. FEEDBACK GAIN PROPERTIES

This section is devoted to determining properties of K(t) to permit calculation of the feedback gains and to establish the nature of the control laws implied by Equations (9) and (10).

The optimal gain matrix K(t), being symmetrical, represents the solution to $\frac{1}{2}(p+1)n[(p+1)n+1]$, nonlinear, ordinary differential equations.

Partitioning K(t) into $(p+1)^2$ $n \times n$ submatrices permits a more detailed examination of Equation (11). Omitting all arguments, the state feedback gain matrix $K_{11}(t)$ is described by

$$\dot{K}_{11} + \left(K_{11}A + \sum_{j=1}^{p} K_{1k} \frac{\partial A}{\partial \alpha_j} \right) + \left(A'K_{11} + \sum_{j=1}^{p} \frac{\partial A'}{\partial \alpha_j} K_{k1} \right) - C_{11} - D_{11} + Q_1(t) = 0 \tag{13}$$

where

$$C_{\mu\nu} = K_{\mu 1} B R_1^{-1} \left(B'K_{1\nu} + \sum_{j=1}^{p} \frac{\partial B'}{\partial \alpha_j} K_{k\nu} \right) \tag{14}$$

and

$$D_{\mu\nu} = \sum_{m=1}^{p} K_{\mu n} \left[M_n K_{n\nu} + \sum_{j=1}^{p} \frac{\partial B}{\partial \alpha_m} R_1^{-1} \frac{\partial B'}{\partial \alpha_j} K_{k\nu} + \frac{\partial B}{\partial \alpha_m} R_1^{-1} B' K_{1\nu} \right] \tag{15}$$

with $k = j + 1$; the indices μ and ν denote row and column, respectively, and $M_n = BR_n^{-1} B'$.

The other submatrices on the diagonal are represented by

$$\dot{K}_{\mu\mu} + K_{\mu\mu}[A + B(J_x u)] + [A + B(J_x u)]' K_{\mu\mu} - C_{\mu\mu} - D_{\mu\mu} + Q_\mu(t) = 0 \qquad (16)$$

where

$$[A + B(J_x u)] = A - BR_1^{-1}\left[B' K_{11} + \sum_{j=1}^{p} \frac{\partial B'}{\partial \alpha_j} K_{k1}\right]. \qquad (17)$$

The submatrices in the first row excluding $K_{11}(t)$ are represented by

$$\dot{K}_{1\nu} + K_{1\nu}[A + B(J_x u)] + \left(A' K_{1\nu} + \sum_{j=1}^{p} \frac{\partial A'}{\partial \alpha_j} K_{k\nu}\right) - C_{1\nu} - D_{1\nu} = 0 \qquad (18)$$

The submatrices in the first column, again excluding $K_{11}(t)$, are given by the transpose of Equation (18). All other submatrices of $K(t)$ are given by

$$\dot{K}_{\mu\nu} + K_{\mu\nu}[A + B(J_x u)] + [A + B(J_x u)]' K_{\mu\nu} - C_{\mu\nu} - D_{\mu\nu} = 0 \qquad (19)$$

It is interesting to compare the preceding equations with the conventional results of optimizing linear systems with respect to a quadratic cost functional. Sensitivity is not accounted for in the standard problem; hence, each submatrix except $K_{11}(t)$ is zero. Thus, Equation (13) reduces to the Riccati equation for the standard optimal control problem.

V. ANALYSIS OF A FIRST-ORDER SYSTEM

In this section, a first-order, time-invariant plant is examined to illustrate the performance which can be obtained with the sensitivity and state controlled configuration. Consider the plant

$$\dot{x}(t) = ax(t) + bu(t), \qquad x(o) = 1 \qquad (20)$$

where the parameter "a" is subject to uncertainty. The cost function is taken to be

$$J = \frac{1}{2}\int_0^{t_f} \left[qx^2(t) + sv^2(t) + r_1 u^2(t) + r_2 m^2(t)\right] dt \qquad (21)$$

The structure of the control configuration is illustrated by Figure 2.

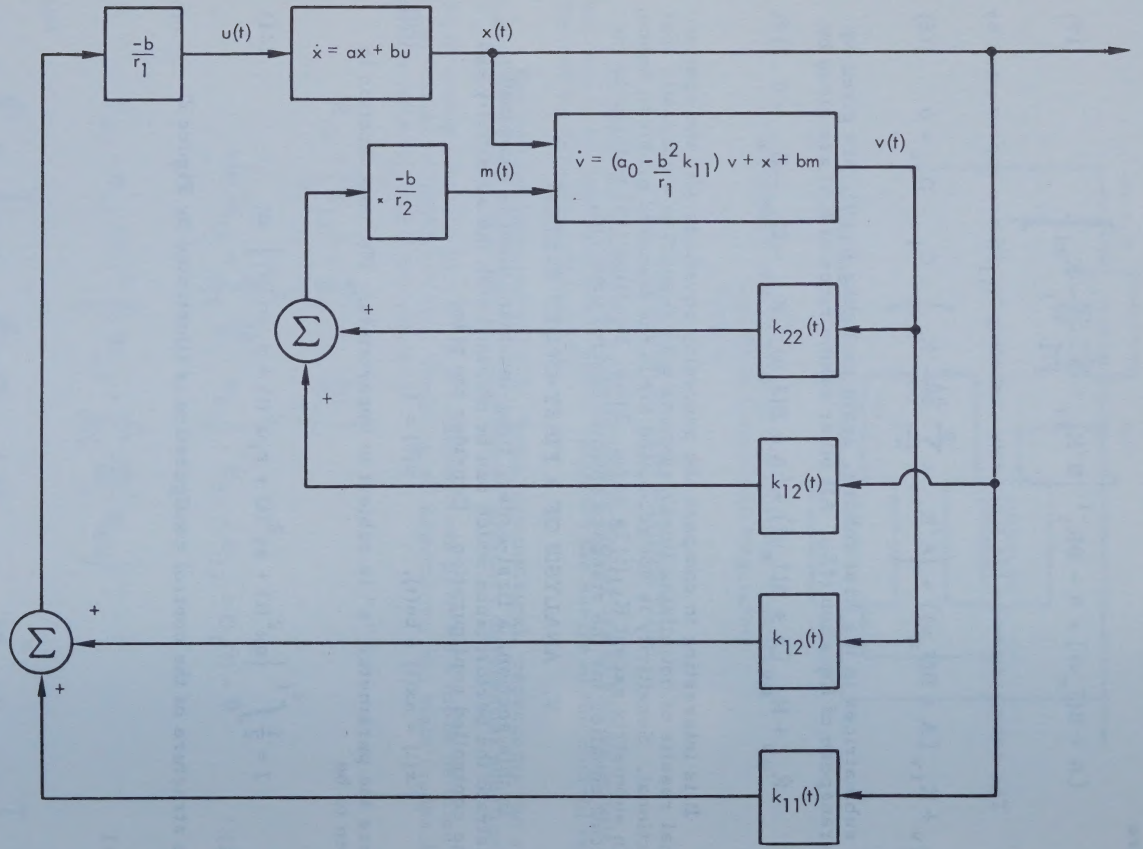


Figure 2. Structure of the Optimal First-Order Sensitivity and State-Controlled Regulating System

For this example, the nominal value of "a" is assumed to be -1.0; but actually a = -0.8. Trade-off data measured in the integral square sense are presented in Figures 3-5 where the coordinates (1, 1) define the error and sensitivity of a plant with no uncertainty, that is, $a = a_o = -0.8$. It may be observed from Figure 3 that when r_2 is small (which means that the cost of sensitivity control is not important), the sensitivity integral square error (ISE) is reduced with relatively little reduction of trajectory error; as r_2 is increased (which means that the cost of sensitivity control becomes important), the state ISE is reduced but sensitivity ISE increases unless s is increased. It might be expected that an increase in state control effort would cause a reduction in sensitivity. This statement is true for a given r_2 if s is increased; however, the contrary is true for a given s with r_2 being increased. The results shown in Figure 3 and 4 are combined in Figure 5 which shows that as r_2 is increased, the cost increases for any given s. This fact indicates that the cost of control effort outweighs the cost of trajectory error.

Some authors (e.g., References 4, 5) have neglected the variation of sensitivity $\partial v/\partial a$. Reference 6 discusses this. By letting $r_2 \to \infty$ (which is approximated by $r_2 = 200$), $m(t) \to 0$; this is equivalent to setting $\partial v/\partial a$ equal to zero. What is the effect of neglecting sensitivity variation? This question can be answered by studying Figure 5. Increasing r_2 (which effectively is applying less and less sensitivity control) for a given s, increases cost. This result points to the improvement in performance to be obtained by not neglecting sensitivity variation in the design.

In Reference 6, the authors were concerned about a sensitivity-controlled trajectory being acceptable. Figure 6 presents the answer for this first-order example. The trajectory maintains the characteristics of an exponential decay except for the case where the sensitivity control signal m(t) is virtually zero.

VI. SUMMARY AND CONCLUSION

In this paper the synthesis of an optimal controller which reduces sensitivity to plant parameter uncertainties has been approached by adding functions of sensitivity and sensitivity variation to a quadratic cost functional. In contrast to a conventional controller, which does not actively account for sensitivity, the sensitivity-constrained controller is more complex but offers a reduction in trajectory error and sensitivity. Implementation of a controller using this approach does not require a parameter tracking scheme (as is often the case in adaptive control designs). Therefore, considering the elimination of a parameter tracking device and the reduction of both trajectory error and sensitivity, any increased controller complexity (that is, added computer memory) may be a justifiable penalty if one has a mandatory sensitivity control requirement.

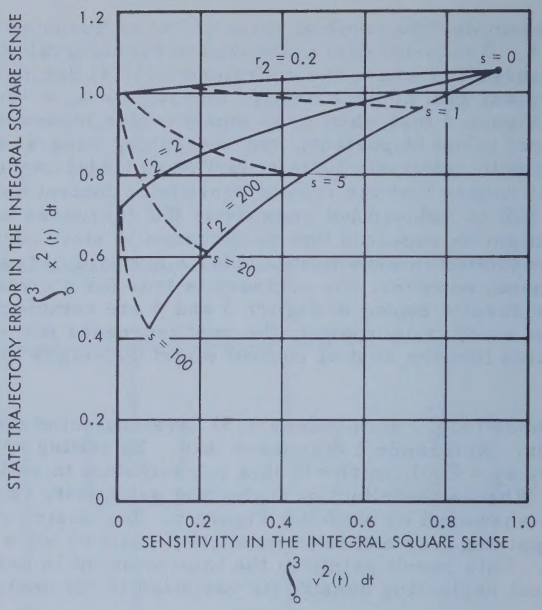


Figure 3. Tradeoff Between Trajectory Error and Sensitivity

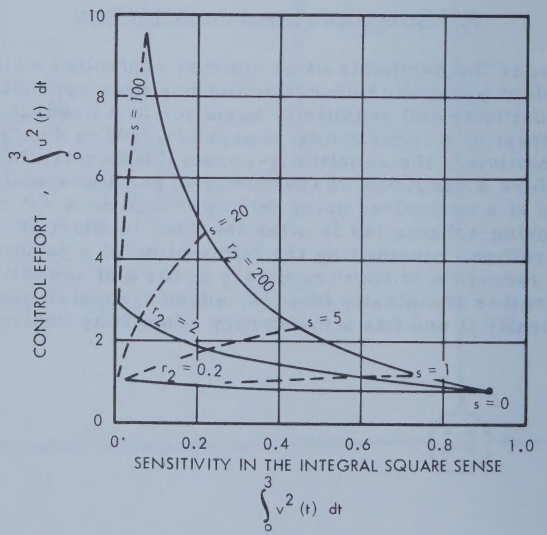


Figure 4. Tradeoff Between Control Effort and Sensitivity

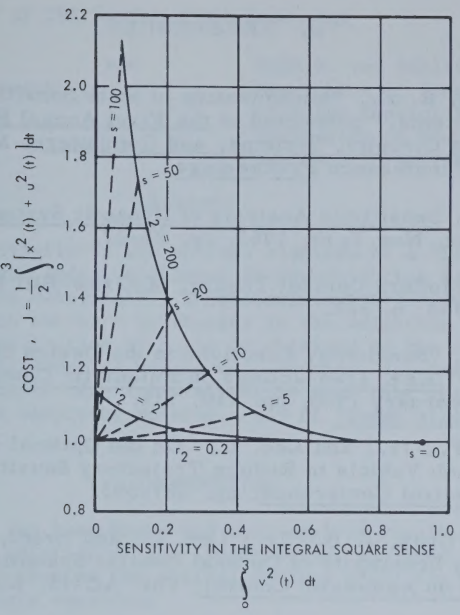


Figure 5. Tradeoff Between Cost and Sensitivity

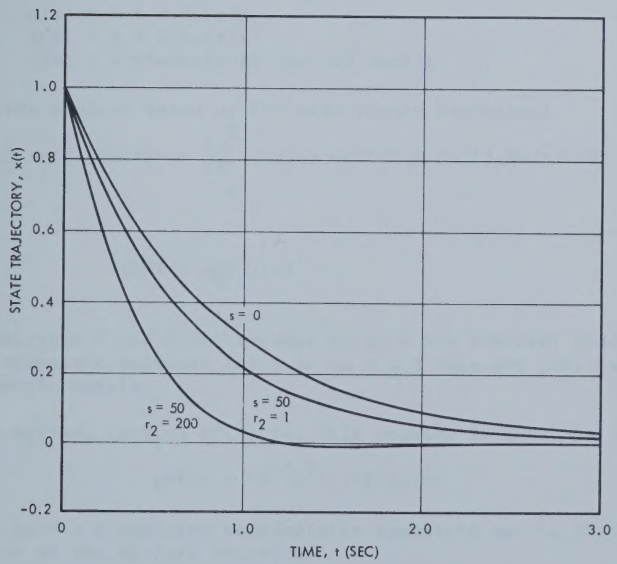


Figure 6. State Trajectory of the Sensitivity-Controlled Regulator System

VII. REFERENCES

1. Higginbotham, R. E., "Minimization of State Sensitivity for Optimal Regulator Systems," presented at the First Annual Houston Conference on Circuits, Systems, and Computers, May 1969. To appear in the conference Proceedings

2. Tomovic, R., Sensitivity Analysis of Dynamic Systems, McGraw-Hill Book Company, New York, 1963, pp. 25-27.

3. Tou, J. T., Modern Control Theory, McGraw-Hill Book Company New York, 1964, p. 222.

4. Bradt, A. J., "Sensitivity Functions in the Design of Optimal Controllers," IEEE Transactions on Automatic Control, Volume AC-13, Number 1, February 1968, pp. 110, 111.

5. Cassidy, J. F., Jr.; and Lee, I., "On the Optimal Feedback Control of a Large Launch Vehicle to Reduce Trajectory Sensitivity," 1967 Joint Automatic Control Conference, pp. 587-595.

6. Sannuti, P.; Cruz, J. B., Jr.; Lee, I.; and Bradt, A. J., "A Note on Trajectory Sensitivity of Optimal Control Systems," IEEE Transactions on Automatic Control, Vol. AC-13, No. 1, February 1968, pp. 111-113.

ON THE SENSITIVITY OF THE LINEAR STATE REGULATOR

DENNIS F. WILKIE
Transportation Research
and Planning Office
Ford Motor Company
Dearborn, Michigan

and

HENK M. van SCHIEVEEN
Department of Computer Information
and Control Engineering
University of Michigan
Ann Arbor, Michigan

ABSTRACT

The sensitivity of the optimal response of a linear system with quadratic performance index to changes in the weighting factors in the performance index is determined. This necessitates finding the sensitivity of the feedback gain matrix K to changes in the weighting factors. It is shown that the sensitivities of K can be obtained as the solution of a set of linear matrix differential equations. Further, for time-invariant systems and an infinite time interval, it is seen that the sensitivities of K are found more simply by solving a set of linear algebraic matrix equations.

I. INTRODUCTION

Probably the best known and most applied result of optimal control theory is the solution of the linear state-regulator problem [1,2]. The problem is to determine the optimal control $\underset{\sim}{u}^*(t)$ for a linear system described by the state equations

$$\dot{\underset{\sim}{x}}(t) = A(t)\underset{\sim}{x}(t) + B(t)\underset{\sim}{u}(t) \ , \ \underset{\sim}{x}(0) = \underset{\sim}{x}^\circ \tag{1}$$

where

$\underset{\sim}{x}(t)$ = n dimensional state vector
$A(t)$ = n x n matrix
$B(t)$ = n x m matrix
$\underset{\sim}{u}(t)$ = m dimensional control vector

which yields the minimum value of the performance functional

$$J(\underset{\sim}{u}) = \frac{1}{2}\langle \underset{\sim}{x}(T), F\underset{\sim}{x}(T)\rangle + \frac{1}{2}\int_0^T \{\langle \underset{\sim}{x}(t), Q\underset{\sim}{x}(t)\rangle + \langle \underset{\sim}{u}(t), R\underset{\sim}{u}(t)\rangle\}dt \ , \tag{2}$$

that is,

$$J(\underset{\sim}{u}^*) = \min_{\underset{\sim}{u}} J(\underset{\sim}{u}) \ . \tag{3}$$

In (2), the matrices F and Q are assumed to be n x n constant positive semidefinite symmetric matrices and R is an m x m constant positive definite symmetric matrix.

The optimal control $\underset{\sim}{u}^*(t)$ for this problem is given in [1,2] as

$$\underset{\sim}{u}^*(t) = -R^{-1}B^T(t)K(t)\underset{\sim}{x}(t) \tag{4}$$

where K(t) is an n x n positive semidefinite symmetric matrix that is the unique solution of the Riccati equation

$$\frac{dK(t)}{dt} = -K(t)A(t) - A^T(t)K(t) + K(t)B(t)R^{-1}B^T(t)K(t) - Q \qquad (5)$$

with the boundary condition $K(T) = F$. The optimal trajectory $\underset{\sim}{x}^*(t)$ of the system (1) with the control law in (4) is then given by the solution of the system of equations

$$\dot{\underset{\sim}{x}}(t) = [A(t) - B(t)R^{-1}B^T(t)K(t)]\underset{\sim}{x}(t) \ , \ \underset{\sim}{x}(0) = \underset{\sim}{x}^\circ \ . \qquad (6)$$

It is evident that the feedback gain matrix $K(t)$, and consequently the optimal trajectory $\underset{\sim}{x}^*(t)$, depend on the values of the elements in the matrices F, Q, and R. Usually, F, Q and R are chosen to be diagonal matrices, †i.e.

$$\begin{aligned} F &= \text{diag } (f_1,\ldots,f_n) \\ Q &= \text{diag } (q_1,\ldots,q_n) \\ R &= \text{diag } (r_1,\ldots,r_m) \ . \end{aligned} \qquad (7)$$

In that case,

$$J(\underset{\sim}{u}) = \frac{1}{2}\sum_{i=1}^{n} f_i x_i^2(T) + \frac{1}{2}\int_0^T \{\sum_{i=1}^{n} q_i x_i^2(t) + \sum_{j=1}^{m} r_j u_j^2(t)\}dt. \qquad (8)$$

Since the weighting factors f_i, q_i, and r_i essentially determine the optimal trajectory, the design aspect of the linear state-regulator problem is determining the appropriate elements of F, Q, and R so that the optimal trajectory is indeed a desireable trajectory.

The choice of a set of weighting factors in (8) that yields suitable system responses may be difficult. For example, consider the problem of longitudinal control of a vehicle which is required to follow a lead vehicle with a desired headway distance and zero relative velocity. This problem is very similar to that considered by Levine and Athans [3]. If the lead vehicle is assumed to be moving at constant velocity, then the deviations of the following car from the desired position and velocity can be modeled by the linearized state equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -ax_2(t) + f(t) \end{aligned} \qquad (9)$$

where

$x_1(t)$ = position error of the following car

$x_2(t)$ = velocity error of the following car

$f(t)$ = control force applied to the following car

a = linearized coefficient of the drag force.

†Note that the results developed in the paper are equally valid for the general case where F, Q, and R are not diagonal.

A reasonable performance functional to use to determine an optimal f(t) in (9) is given in [3] as

$$J(f) = \frac{1}{2}\int_0^\infty \{q_1 x_1^2(t) + q_2 x_2^2(t) + r_1 f^2(t)\}dt \qquad (10)$$

where q_1, q_2, and r_1 are positive, constant weighting factors which must be specified. Levine and Athans[3] have obtained good results for relatively small initial conditions for $x_1(t)$ and $x_2(t)$ which represent deviations of the system state from a desired operating point. This problem corresponds to that of station keeping once the vehicles have reached a desired region of the nominal operating point. In the vehicle control problem associated with automated highways, initial position errors may be considered to vary from 0-500 feet while initial velocity errors may be as high as 50 miles per hour. In such a case, it is difficult to determine weighting factors which will yield acceptable vehicle responses over the whole range of expected initial conditions. It would be very desireable if some technique, other than a random search technique, for determining a suitable set of weighting factors was available. It is apparent that a means of investigating the effects of changes in weighting factors on the optimal system response would be helpful in this regard.

The problem of relating changes in the weighting factors to changes in the eigenvalues of the optimal closed-loop system matrix $A(t)-B(t)R^{-1}B^T(t)K(t)$ has been considered by Chang [4], Rynaski [5], Tyler and Tuteur [6], and Chen and Shen [7]. Chang originated the root square locus to investigate the effects of a weighting factor on the closed loop eigenvalues for single-input single-output systems, and Rynaski, Tyler and Tuteur extended these results to multi-variable systems. Chen and Shen derived a relationship between changes in the weighting factors and changes in the eigenvalues of the closed-loop system. Neither of these techniques directly relates the sensitivity of the optimal system response to changes in the weighting factors. Thus, if the designer is concerned with obtaining a "suitable optimal trajectory," it would be necessary to relate changes in trajectory to changes in the eigenvalues of the closed-loop system matrix.

In this paper, the sensitivity equations for the state vector of the system (1) with respect to changes in the weighting factors are developed. These equations require the determination of the sensitivities of the feedback gain matrix K(t) to changes in the weighting factors, i.e. $\partial K/\partial f_i$, $\partial K/\partial q_i$, i=1,...,n and $\partial K/\partial r_j$, j=1,...,m, which can be obtained by solving a set of <u>linear</u> matrix differential equations.

The results are particularly useful in the case of time-invariant systems with $T=\infty$. In that case $\partial K/\partial f_i$, $\partial K/\partial q_i$ and $\partial K/\partial r_j$ (i=1,...n, j=1,...,m) can be obtained by solving a set of <u>linear algebraic</u> equations. Because of this result, the sensitivities of the optimal trajectory to changes in the weighting factors are quite easily obtained.

The use of these techniques in a design procedure involving either trial and error or perhaps two-level design is discussed in section IV.

Finally, the techniques of the paper are used to determine an appropriate set of weighting factors for a vehicle-following control system.

II. THE SENSITIVITY EQUATIONS

The problem to be considered is precisely stated as follows:

Given the linear system

$$\dot{\underset{\sim}{x}}(t) = A(t)\underset{\sim}{x}(t) + B(t)\underset{\sim}{u}(t) \ , \quad \underset{\sim}{x}(0) = \underset{\sim}{x}^{\circ} \tag{11}$$

and the performance functional

$$J(\underset{\sim}{u}) = \frac{1}{2}\langle \underset{\sim}{x}(T), F\underset{\sim}{x}(T)\rangle + \frac{1}{2}\int_{0}^{T} \{\langle \underset{\sim}{x}(t), Q\underset{\sim}{x}(t)\rangle + \langle \underset{\sim}{u}(t), R\underset{\sim}{u}(t)\rangle\}dt \tag{12}$$

where F, Q, and R are diagonal constant matrices as in (7), determine the sensitivity of the optimal trajectory $\underset{\sim}{x}^*(t)$ to variations in the weighting factors $f_i \varepsilon F$, $q_i \varepsilon Q$, or $r_j \varepsilon R$. That is, determine $\partial\underset{\sim}{x}^*/\partial f_i$, $\partial\underset{\sim}{x}^*/\partial q_i$, i=1,...,n and $\partial\underset{\sim}{x}^*/\partial r_j$, j=1,...,m.†

The sensitivity equations which must be solved to obtain $\partial\underset{\sim}{x}^*/\partial f_i$, $\partial\underset{\sim}{x}^*/\partial q_i$ and $\partial\underset{\sim}{x}^*/\partial r_j$ can be obtained by differentiating (6) with respect to the parameters f_i, q_i, and r_j. Since there are differences in the sensitivity equations for the various parameters, the three cases are considered separately.

Case (A), $\partial\underset{\sim}{x}^*(t)/\partial f_i$, i=1,...n:

Differentiation of (6) with respect to f_i yields the vector differential equation

$$\frac{\partial}{\partial f_i}(\dot{\underset{\sim}{x}}^*(t)) = \frac{d}{dt}\left[\frac{\partial\underset{\sim}{x}^*(t)}{\partial f_i}\right] = [A(t) - B(t)R^{-1}B^T(t)K(t)]\frac{\partial\underset{\sim}{x}^*(t)}{\partial f_i}$$

$$- B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial f_i}\underset{\sim}{x}^*(t), \quad \left.\frac{\partial\underset{\sim}{x}^*}{\partial f_i}\right|_{t=0} = \underset{\sim}{0}, \tag{13}$$

since the necessary conditions for interchanging the order of differentiation are satisfied. The initial condition for the sensitivity equation is as shown since the initial state $\underset{\sim}{x}^\circ$ is independent of the weighting factors.

The solution of (13) requires knowledge of the sensitivity of the feedback gain matrix K(t) with respect to changes in f_i. The equation which determines this sensitivity can be obtained by differentiation of (5) with respect to f_i which yields the linear matrix differential equation

†Since sensitivity functions are always considered to be evaluated at some specific parameter value, $\partial\underset{\sim}{x}^*/\partial f_i$ will be used to represent $\left.\frac{\partial\underset{\sim}{x}^*}{\partial f_i}\right|_{F^\circ,Q^\circ,R^\circ}$ and all other partial derivatives will be treated similarly.

$$\frac{d}{dt}\left[\frac{\partial K(t)}{\partial f_i}\right] = \frac{-\partial K(t)}{\partial f_i}A(t) - A^T(t)\frac{\partial K(t)}{\partial f_i} + \frac{\partial K(t)}{\partial f_i}B(t)R^{-1}B^T(t)K(t)$$

$$+ K(t)B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial f_i} . \qquad (14)$$

The boundary condition for (14) is obtained by using the fact that $K(T) = F$ and thus

$$\left.\frac{\partial K(t)}{\partial f_i}\right|_{t=T} = \text{diag}\ (\delta_{i1}, \delta_{i2}, \ldots, \delta_{in}) , \qquad (15)$$

where δ_{ij} is the Kronecker δ.

Case (B), $\partial \underset{\sim}{x}^*(t)/\partial q_i$, i=1,...,n:

Differentiation of (6) with respect to q_i yields the vector differential equation

$$\frac{d}{dt}\left[\frac{\partial \underset{\sim}{x}^*(t)}{\partial q_i}\right] = [A(t) - B(t)R^{-1}B^T(t)K(t)]\frac{\partial \underset{\sim}{x}^*(t)}{\partial q_i}$$

$$-B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial q_i}\underset{\sim}{x}^*(t) \ , \ \left.\frac{\partial \underset{\sim}{x}^*}{\partial q_i}\right|_{t=0} = \underset{\sim}{0}. \qquad (16)$$

The differential equation for $\partial K(t)/\partial q_i$ is obtained by differentiating (5) with respect to q_i which yields the linear matrix differential equation

$$\frac{d}{dt}\left[\frac{\partial K(t)}{\partial q_i}\right] = \frac{-\partial K(t)}{\partial q_i}A(t) - A^T(t)\frac{\partial K(t)}{\partial q_i} + \frac{\partial K(t)}{\partial q_i}B(t)R^{-1}B^T(t)K(t)$$

$$+ K(t)B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial q_i} - \frac{\partial Q}{\partial q_i} , \qquad (17)$$

where for Q as in (7),

$$\frac{\partial Q}{\partial q_i} = \text{diag}\ (\delta_{i1}, \delta_{i2}, \ldots, \delta_{in}) . \qquad (18)$$

The boundary condition for (17) is $\left.\frac{\partial K(t)}{\partial q_i}\right|_{t=T} = \frac{\partial F}{\partial q_i} = [0]$.

Case (C), $\partial \underset{\sim}{x}^*(t)/\partial r_j$, j=1,...,m:

Differentiation of (6) with respect to r_j yields the vector differential equation

$$\frac{d}{dt}\left[\frac{\partial \underset{\sim}{x}^*(t)}{\partial r_j}\right] = [A(t) - B(t)R^{-1}B^T(t)K(t)]\frac{\partial \underset{\sim}{x}^*(t)}{\partial r_j}$$

$$- [B(t)\frac{\partial(R^{-1})}{\partial r_j}B^T(t)K(t) + B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial r_j}]\underset{\sim}{x}^*(t), \frac{\partial \underset{\sim}{x}^*}{\partial r_j}_{t=0} = \underset{\sim}{0} . \quad (19)$$

For the chosen form of R in (7),

$$\frac{\partial(R^{-1})}{\partial r_j} = \frac{\partial}{\partial r_j}[\text{diag}\ (1/r_1, 1/r_2, \ldots, 1/r_m)]$$

$$= -\text{diag}\left(\frac{\delta_{j1}}{r_1^2}, \frac{\delta_{j2}}{r_2^2}, \ldots, \frac{\delta_{jm}}{r_m^2}\right)$$

As in the previous cases, the linear matrix differential equation for $\partial K(t)/\partial r_j$ is found by differentiation of (5) with respect to r_j, which yields

$$\frac{d}{dt}\left[\frac{\partial K(t)}{\partial r_j}\right] = -\frac{\partial K(t)}{\partial r_j}A(t) - A^T(t)\frac{\partial K(t)}{\partial r_j}$$

$$+ \frac{\partial K(t)}{\partial r_j}B(t)R^{-1}B^T(t)K(t) + K(t)B(t)\frac{\partial(R^{-1})}{\partial r_j}B^T(t)K(t)$$

$$+ K(t)B(t)R^{-1}B^T(t)\frac{\partial K(t)}{\partial r_j} , \quad (21)$$

with the boundary condition $\left.\frac{\partial K(t)}{\partial r_j}\right|_{t=T} = [0]$.

Thus, the sensitivity functions $\partial\underset{\sim}{x}^*/\partial f_i$, $\partial\underset{\sim}{x}^*/\partial q_i$ and $\partial\underset{\sim}{x}^*/\partial r_j$ i=1,...,n,j=1,...,m can be found by solving (13) and (14), (16) and (17), or (19) and (21) respectively. Since the matrix K(t) is symmetric, the solution of (14), (17) or (21) requires the solution of n(n+1)/2 first-order differential equations for each $\partial K(t)/\partial f_i$, $\partial K(t)/\partial q_i$, or $\partial K(t)/\partial r_j$ desired. Then, to determine the sensitivity functions $\partial\underset{\sim}{x}^*/\partial f_i$, $\partial\underset{\sim}{x}^*/\partial q_i$ and $\partial\underset{\sim}{x}^*/\partial r_j$ it is necessary to solve (13), (16), and (19) respectively, each of which involves the solution of n first-order differential equations. Therefore, the number of differential equations that must be solved for a given set of weighting factors to determine the optimal trajectory and <u>all</u> the sensitivity functions of the optimal trajectory with respect to the 2n + m weighting factors is

$$N = (1 + 2n + m)(n + \frac{n(n+1)}{2}) . \quad (22)$$

This may be an unacceptable number of differential equations, even for low order systems. Often not all the sensitivities of the optimal trajectory with respect to changes in the weighting factors would have to be determined, and the number of differential equations to be solved would thus be smaller.

Essential simplifications can be made in the procedure for determining the sensitivity of the optimal trajectory to changes in weighting factors in the case of time-invariant systems with $T = \infty$. These results are discussed in the next section.

III. TIME-INVARIANT SYSTEMS WITH T= ∞

For time-invariant systems with T= ∞, the state equation of the system becomes

$$\dot{\underset{\sim}{x}}(t) = A\underset{\sim}{x}(t) + B\underset{\sim}{u}(t) \tag{23}$$

and the performance functional is

$$J(\underset{\sim}{u}) = \frac{1}{2}\int_0^\infty \{<\underset{\sim}{x}(t), Q\underset{\sim}{x}(t)> + <\underset{\sim}{u}(t), R\underset{\sim}{u}(t)>\}dt.$$

If the system (23) is controllable and Q is positive definite (rather than semidefinite as in the time-varying case) then the optimal control $\underset{\sim}{u}^*(t)$ which minimizes $J(\underset{\sim}{u})$ as given in [1] is

$$\underset{\sim}{u}^*(t) = R^{-1}B^T\hat{K}\underset{\sim}{x}(t) \tag{24}$$

where $\hat{K}$ is a constant positive definite matrix which satisfies the non-linear algebraic equation

$$-\hat{K}A - A^T\hat{K} + \hat{K}BR^{-1}B^T\hat{K} - Q = [0]. \tag{25}$$

Alternatively $\hat{K}$ can be found by solving the Riccati equation (5) for K(t) with T→ ∞ since as shown by Kalman [2]

$$\hat{K} = \lim_{T\to\infty} K(t). \tag{26}$$

As in the previous case, $\hat{K}$ depends on the elements of Q and R. From the results of the previous section, $\partial\underset{\sim}{x}^*(t)/\partial q_i$ and $\partial\underset{\sim}{x}^*(t)/\partial r_j$, i=1,...,n,j=1,...m are given by

$$\frac{d}{dt}\left[\frac{\partial\underset{\sim}{x}^*(t)}{\partial q_i}\right] = [A - BR^{-1}B^T\hat{K}]\frac{\partial\underset{\sim}{x}^*(t)}{\partial q_i} - BR^{-1}B^T\frac{\partial\hat{K}}{\partial q_i}\underset{\sim}{x}^*(t) \tag{27}$$

and

$$\frac{d}{dt}\left[\frac{\partial\underset{\sim}{x}^*(t)}{\partial r_j}\right] = [A - BR^{-1}B^T\hat{K}]\frac{\partial\underset{\sim}{x}^*(t)}{\partial r_j} + [B\frac{\partial R^{-1}}{\partial r_j}B^T\hat{K} - BR^{-1}B^T\frac{\partial K}{\partial r_j}]\underset{\sim}{x}^*(t)$$

Differentiating (25) with respect to q_i or r_j and rearranging the terms yields the following equations for $\partial\hat{K}/\partial q_i$ and $\partial\hat{K}/\partial r_j$,

$$\frac{\partial\hat{K}}{\partial q_i}(A - BR^{-1}B^T\hat{K}) + (A^T - \hat{K}BR^{-1}B^T)\frac{\partial\hat{K}}{\partial q_i} = -\frac{\partial Q}{\partial q_i}, \quad i=1,\ldots,n \tag{29}$$

and

$$\frac{\partial\hat{K}}{\partial r_j}(A - BR^{-1}B^T\hat{K}) + (A^T - \hat{K}BR^{-1}B^T)\frac{\partial\hat{K}}{\partial r_j} = \hat{K}B\frac{\partial(R^{-1})}{\partial r_j}B^T\hat{K}, \quad j=1,\ldots m. \tag{30}$$

These linear algebraic matrix equations have the form of the Lyapunov matrix equation which arises in the stability analysis and the calculation of performance indices for linear systems [8-10]. An algorithm has been given for the solution of the Lyapunov equation [9] which requires much less computation time than that required for the solution of the differential equations (17) and (21) in the time-varying case.

Consider the number of differential equations which must be solved in this case. In addition to the n differential equations which determine $\underset{\sim}{x}^*(t)$ and the $n(n+1)/2$ differential equations to compute $\hat{K}$, the n sensitivity equations must be solved for each q_i or r_j which is considered (equations (20) or (30)). Thus, at most $N = n(n+1)/2 + i + n + m$ differential equations must be solved, which is considerably less than the number that must be solved in the time-varying case.

In the case of single input systems, further computational savings could be realized by using the techniques given in [11]. Using those techniques, the sensitivity functions $\partial \underset{\sim}{x}^*/\partial k_{ij}$, where $\hat{K} = [k_{ij}]$, could be determined using one dynamic nth order sensitivty model in addition to the system, i,e., only 2n differential equations would have to be solved. Then, using $\partial \hat{K}/\partial q_i$ and $\partial \hat{K}/\partial r_j$ as obtained from (29) and (30), the sensitivity of the optimal trajectory to changes in weighting factors would be found to be

$$\frac{\partial \underset{\sim}{x}^*}{\partial q_i} = \sum_{\ell=1}^{n} \sum_{m=1}^{n} \left(\frac{\partial \underset{\sim}{x}^*}{\partial k_{\ell m}} \right) \left(\frac{\partial k_{\ell m}}{\partial r_j} \right) \tag{31}$$

and

$$\frac{\partial \underset{\sim}{x}^*}{\partial f_j} \sum_{\ell=1}^{n} \sum_{m=1}^{n} \left(\frac{\partial \underset{\sim}{x}^*}{\partial k_{\ell m}} \right) \left(\frac{\partial k_{\ell m}}{\partial r_j} \right) \tag{32}$$

IV. USE OF THE SENSITIVITY FUNCTIONS IN DESIGN PROBLEMS

Since, as discussed in the introduction, the choice of the weighting factors in the performance functional in (2) is often difficult and initially quite arbitrary, an iterative technique for determining suitable weighting factors using the results of the previous sections is proposed. The procedure consists of the following steps:

(1) Choose an initial set of weighting factors and compute the optimal control and trajectory for the assumed weighting factors.

(2) Determine whether the present set of weighting factors yields a control which produces acceptable responses. If the responses are acceptable, the problem is solved, otherwise proceed with step 3.

(3) Determine the sensitivity functions $\partial \underset{\sim}{x}^*/\partial q_i$, $\partial \underset{\sim}{x}^*/\partial f_i$, and $\partial \underset{\sim}{x}^*/\partial r_j$, i=1,..., n, j=1,...,m from the information obtained in step (1) and the results of the previous sections. For notational convenience, let the present values of the weighting factors be denoted as follows,

$$\underset{\sim}{f}^\circ = \begin{bmatrix} f_1^\circ \\ \cdot \\ \cdot \\ \cdot \\ f_n^\circ \end{bmatrix}, \quad \underset{\sim}{q}^\circ = \begin{bmatrix} q_1^\circ \\ \cdot \\ \cdot \\ \cdot \\ q_n^\circ \end{bmatrix} \text{ and } \underset{\sim}{r}^\circ = \begin{bmatrix} r_1^\circ \\ \cdot \\ \cdot \\ \cdot \\ r_m^\circ \end{bmatrix}.$$

Expanding $\underset{\sim}{x}^*(t, \underset{\sim}{f}, \underset{\sim}{q}, \underset{\sim}{r})$ in a Taylor series about the nominal point $\underset{\sim}{f}^\circ$, $\underset{\sim}{q}^\circ$, $\underset{\sim}{r}^\circ$ gives

$$\underset{\sim}{x}^*(t,\underset{\sim}{f},\underset{\sim}{q},\underset{\sim}{r}) = \underset{\sim}{x}^*(t,\underset{\sim}{f}^\circ,\underset{\sim}{q}^\circ,\underset{\sim}{r}^\circ) + \left.\frac{\partial \underset{\sim}{x}^*}{\partial \underset{\sim}{f}}\right|_o (\underset{\sim}{f}-\underset{\sim}{f}^\circ)$$

$$+ \left.\frac{\partial \underset{\sim}{x}^*}{\partial \underset{\sim}{q}}\right|_o (\underset{\sim}{q}-\underset{\sim}{q}^\circ) + \left.\frac{\partial \underset{\sim}{x}^*}{\partial \underset{\sim}{r}}\right|_o (\underset{\sim}{r}-\underset{\sim}{r}^\circ) + \text{higher-order terms} \qquad (33)$$

where $\left.\partial \underset{\sim}{x}^*/\partial \underset{\sim}{f}\right|_o$, $\left.\partial \underset{\sim}{x}^*/\partial \underset{\sim}{q}\right|_o$, $\left.\partial \underset{\sim}{x}^*/\partial \underset{\sim}{r}\right|_o$ are Jacobian matrices evaluated at $\underset{\sim}{f}^\circ, \underset{\sim}{q}^\circ, \underset{\sim}{r}^\circ$.

Now by inspection of the unacceptable components of the state trajectory and the sensitivity functions $\partial \underset{\sim}{x}^*/\partial \underset{\sim}{f}$, $\partial \underset{\sim}{x}^*/\partial \underset{\sim}{q}$ and $\partial \underset{\sim}{x}^*/\partial \underset{\sim}{r}$, a small change in the weighting factors can be determined from (33) which will improve the system response.

Steps (1) through (3) are repeated iteratively until a suitable set of weighting factors and control have been determined.

This design procedure is certainly a heuristic one, and no guarantee can be made that a suitable design will be achieved. However, it is a better procedure for determining the best set of weighting factors than simply guessing, since it is guessing with increased information.

Alternatively, the sensitivity functions $\partial \underset{\sim}{x}^*/\partial \underset{\sim}{f}, \partial \underset{\sim}{x}^*/\partial q$ and $\partial \underset{\sim}{x}^*/\partial \underset{\sim}{r}$, could perhaps be used with numerical techniques in a two-level design approach. For example, the behavior of a system might be considered "generally optimal" if a performance functional of the form

$$J_1 = \frac{1}{2}\langle \underset{\sim}{x}(T), F\underset{\sim}{x}(T)\rangle + \frac{1}{2}\int_o^T \{\langle \underset{\sim}{x}(t), Q\underset{\sim}{x}(t)\rangle + \langle \underset{\sim}{u}(t), R\underset{\sim}{u}(t)\rangle\}dt \qquad (34)$$

is minimized. Suppose that it is difficult to obtain a trajectory which is optimal for a given set of weighting factors in J_1 and also yields acceptable responses for a subset $\underset{\sim}{x}_C^*$ of the state variables. Perhaps a second performance functional J_2 could be defined

$$J_2 = \int_o^T L(\underset{\sim}{x}_C^*)dt \qquad (35)$$

which is a measure of the unacceptibility of the response of the subset $\underset{\sim}{x}_C^*$ and therefore should be minimized. Obviously J_2 is a function of the weighting vectors $\underset{\sim}{f}$, $\underset{\sim}{q}$ and $\underset{\sim}{r}$. Then the two-level design procedure would involve using an iterative numerical procedure to choose the weighting vectors $\underset{\sim}{f}^*$, $\underset{\sim}{q}^*$ and $\underset{\sim}{r}^*$ so that

$$J_2(\underset{\sim}{f}^*, \underset{\sim}{q}^*, \underset{\sim}{r}^*) = \min_{\underset{\sim}{f},\underset{\sim}{q},\underset{\sim}{r}} \int_o^T L(\underset{\sim}{x}_C^*)dt, \qquad (36)$$

where $\underset{\sim}{y}(t)$ would be chosen to minimize J_1 for the given weighting factors. If a gradient type method was used in the minimization indicated in (36), then the gradient components

$$\frac{\partial J_2}{\partial w} = \int_0^T \operatorname{grad}_{\underset{\sim}{x}_C^*} L(x_C^*) \frac{\partial \underset{\sim}{x}_C^*}{\partial w} \, dt \tag{37}$$

where w, $\underset{\sim}{f}$, q, or $\underset{\sim}{r}$ could be obtained using the results of the paper to generate $\partial \underset{\sim}{x}_C^*/\partial w$.

Essentially this two-level design procedure is an attempt to give the procedure outlined above in steps (1)-(3) a more rigorous mathematical meaning. However, there may be difficulties in choosing the appropriate performance functional J_2.

V. EXAMPLE

As an example of applying the results of the paper, consider the problem of determining a suitable set of weighting factors in the longitudinal vehicle control problem discussed earlier. Suppose that the controlled vehicle is to maintain a given position on the roadway (which is defined by a control policy) as well as maintaining a nominal system velocity in the steady state. Using a third-order model instead of the second-order model discussed earlier, the deviations of the vehicle position, velocity, and acceleration from the desired values can be modeled by the linearized state equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= x_3(t) \\ \dot{x}_3(t) &= -\bar{a}x_3(t) + w(t) \end{aligned} \tag{38}$$

where

$x_1(t)$ = position error of the controlled vehicle
$x_2(t)$ = velocity error of the controlled vehicle
$x_3(t)$ = acceleration of the controlled vehicle,

and $w(t)$ is to be chosen so that the performance functional

$$J(w) = \frac{1}{2} \int_0^\infty \{\langle \underset{\sim}{x}, Q\underset{\sim}{x} \rangle + rw^2\} dt$$

is minimized.

Stated in this manner, the vehicle control problem corresponds to following hypothetical slots rather than car following. In this approach, it is assumed that a vehicle has attained steady state (zero-error) position, velocity, and acceleration when it receives a command to change to a new position with the same nominal velocity and zero acceleration in the steady state. Thus, the initial position error is the only non-zero initial condition for the error state equations. The problem is to determine the diagonal elements of the Q matrix and R so that the trajectories obtained for the expected range of initial position errors satisfy the conditions that

(a) the ride is comfortable
(b) it does not take too long to achieve the desired final position
(c) car capabilities are not exceeded.

It is apparent that (a) and (c) essentially represent constraints on the acceleration profile of the vehicle.

The trajectory for an initial position error of 100 feet and weighting factors $q_i=1$, $i=1,\ldots,3$, $r=1$ is shown in Figure 1. It is seen that the maximum acceleration is about 24 ft/sec^2 or .75g. This is too great an acceleration to expect from the vehicle and also it would not represent a comfortable ride. The sensitivities of the acceleration profile to changes in the various weighting factors are shown in Figure 2. It is apparent from these that one way to decrease the magnitude of the peaks of the acceleration profile would be to decrease w_1. However, the other weighting factors also affect the response of the system and could be adjusted accordingly to help obtain a better acceleration curve. Using the heuristic design approach discussed in the previous section, weighting factors were obtained which yield the system response shown in Figure 3 which is indeed a much more acceptable response. The vehicle is seen to be within one foot of its final position in 18 seconds while experiencing a peak acceleration of only 2.9 ft/sec^2 , and this is acceptable in terms of the above performance criteria.

VII. REFERENCES

1. M. Athans and P. L. Falb, Optimal Control, McGraw Hill, New York, 1966, Chapter 9.

2. R. E. Kalman, "When is a Linear System Optimal?", Trans. ASME J. Basic Engineering, March, 1964, pp. 51-60.

3. W. S. Levine and M. Athans, "On the Optimal Error Regulation of a String of Moving Vehicles," IEEE Trans. Automatic Control, Vol. AC-11, No. 3, July, 1966, pp. 355-361.

4. S. S. L. Chang, Synthesis of Optimum Control Systems, New York, McGraw Hill, 1961.

5. E. G. Rynaski, "The Multivariable Root-Square Locus -- An Optimal System Design Aid," AIAA Conference on Guidance and Control, Minneapolis, Minnesota, 1965.

6. J. S. Tyler and F. B. Tuteur, "The Use of a Quadratic Performance Index to Design Multivariable Systems," IEEE Trans. Automatic Control, vol. AC-11, No. 1, January, 1966, pp. 84-92.

7. R. T. N. Chen and D. W. C. Shen, "Sensitivity Analysis and Design of Multivariable Regulators Using a Quadratic Performance Criterion," Preprints Ninth JACC: June, 1968; University of Michigan, Ann Arbor, Michigan, pp. 229-237.

8. R. E. Kalman and J. E. Bertram, "Control System Analysis and Design Via the Second Method of Lyapunov," Trans. ASME, J. Basic Eng., vol. 82, Series D, June 1960, pp. 371-400.

9. E. J. Davison and F. T. Mann, "The Numerical Solution of $A^TQ + QA = -C$," IEEE Trans. Automatic Control, Vol. AC-13, August, 1968, pp. 448-449.

10. S. Barnett and C. Storey, "The Liapunov Matrix Equation and Schwarz's Form," IEEE Trans. Automatic Control, vol. AC-12, February, 1967, pp. 117-118.

11. D. F. Wilkie and W. R. Perkins, "Essential Parameters in Sensitivity Analysis," Automatica, Vol. 5, No. 2, March, 1969, pp. 191-198.

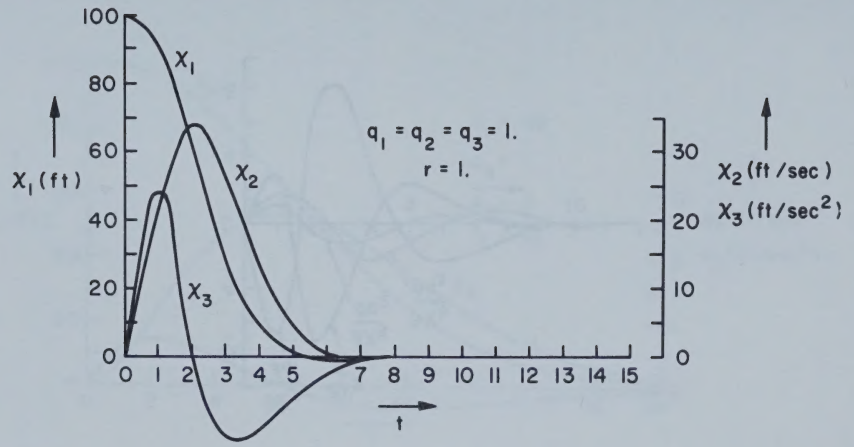


Figure 3. System Trajectory for Final Weighting Factors.

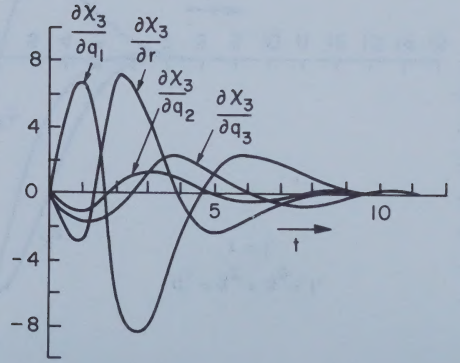


Figure 2. Sensitivity Functions with respect to Weighting Factors.

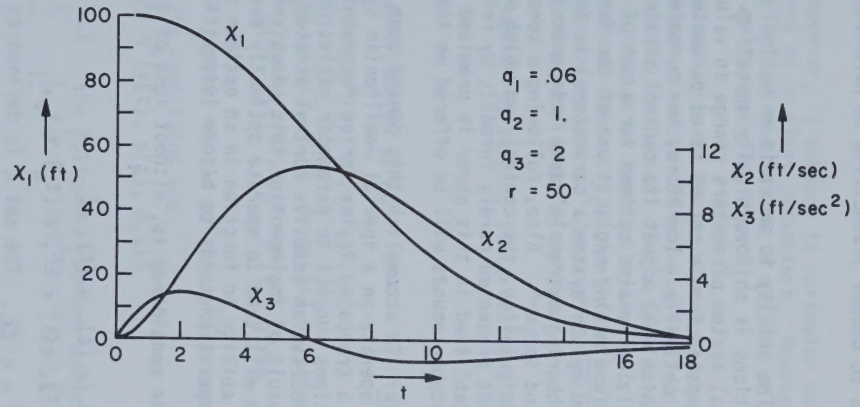


Figure 1. System Trajectory for Initial Set of Weighting Factors.

THE DESIGN OF AN OPTIMAL SYSTEM SUBJECT TO PARAMETER VARIATIONS

FRED J. TAYLOR
Member of The Technical Staff
Texas Instruments, Dallas, Texas

Abstract: In this paper control law is derived which will continually adjust its control strategy in such a manner that the control remains optinum for a system subject to parameter variations. To accomplish this the Maximum Principal will operate on a special Taylor series representation of a Hamiltonian system of equations. As a result of this optimization process, the system tends to become insensitive to parameter variations.

Introduction: The ability to maintain an optimal solution independent of parameter variations is philosophically appealing. Unfortunately, whenever the internal system parameters change in value, and the control policy has been pre-computed from a set of fixed parameters, the system usually no longer operates optimally. Therefore, the purpose of this paper is to devise a scheme which will adjust its control strategy in such a manner so that the control law remains optimum for a set of parameter variations. This result (before making explicit use of the fact that the state vector can be agumented by the system's parameters) is formally similar to the theory of "neighboring extremals" done independently by Kelly[1] and by Breakwell, Bryson, and Speyer[2]. Also, following a considerable amount of algebraic matrix manipulation, the control law which will be developed could be derived from that presented purely formally by Friedland[3]. However, the control law established in this paper is premised on more rigorous arguments and computable bounds will be offered on the solutions domain of validity.

Method of Attack: To accomplish this demand upon the system, the Maximum Principle will operate on a special Hamiltonian system. The special system referred to is a truncated Taylor series representation of a Hamiltonian system of equations subject to parameter variations. As a result of the optimization process, an adaptive control strategy will evolve as a derived result. The result of implemention this adaptiveness into the system is that the system will tend to operate optimally over a set of parameter variations. This ability to function in an optimum mode produces a system whose cost of operation tends to become insensitive to parameter variations.

The system to be considered is, without loss of generality, the autonomous plant:

$$\dot{x}(t) = f(x(t), u(t); \alpha) \tag{1}$$

where $t \in [t_0, T]$, $x(t) \in E^n$, $x(t_0) = x_0$, $u(t) \in \Omega$, and $\alpha \in \mathcal{A}$. The set Ω is defined to be:

$$\Omega \triangleq \left\{ u(t) = \begin{bmatrix} u_1(t) \\ \\ u_r(t) \end{bmatrix} \;\middle|\; \begin{array}{l} u_i(t) \text{ is an admissible measurable} \\ \text{control function on } t \in [t_0, T], \\ i = 1, \ldots, r. \end{array} \right\}$$

and $\mathcal{A} = \left\{ \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} \;\middle|\; \alpha_i \text{ is an admissible measurable parameter of system (1)} \right\}$

The performance index C(u) will be defined to be:

$$C(u) = 1/2 < x(t), \Gamma x(t) > + \int_{t_0}^{T} L(x(t), u(t))\, dt, \quad \|\Gamma\| \geq 0. \qquad (2)$$

Necessary conditions, for a fixed parameter vector α, for finding a u (t) which transfers the state of system (1) from x_0 to some target set S at the fixed terminal time T such that C(u) is minimized, subject to the differential side constraint (1), are well known from the Maximum Principle[4].

The system under study however, is subject to parameter variations. These parameters will be defined to be plant parameters. However, equivalent alternative parameters are initial conditions, a mixture of plant parameters and initial ocnditions, or the starting time t_0.

This study of the optimal solution to the problem subject to parameter variations will now consider extension of known optimal solutions into a neighborhood of a known optimal solution. The extension mechanixm to be investigated will be an expansion of the Hamiltonian by a Truncated Taylor Series about the assumed known nominally optimal solution (i.e., the optimal solution which corresponds to $\alpha = \alpha^n$ where α^n is the nominal design parameter vector). This is accomplished as follows:

Define a new state vector (augmented state vector) y such that

$y \in E^n \times \mathcal{A} \subset E^{n+m}$

where $y \triangleq \begin{bmatrix} x \\ \alpha \end{bmatrix}$

Define the nominal Hamiltonian to be:

$$H^n(y(t), p(t), u(t)) = H(y^n(t), p^n(t), u^n(t))$$

for $p(t) \in E^{n+m}$ and $y^n(t)$, $p^n(t)$, $u^n(t)$ to be the solutions of

$$\dot{y}^n(t) = \partial H^n(y(t), p(t), u(t)) / \partial p^n, \; y^n(t_0) \text{ given} \qquad (4)$$

$$\dot{p}^n(t) = -\partial H^n(y(t), p(t), u(t)) / \partial y^n(t), \; p^n(T) \text{ known from the usual transversality conditions and} \qquad (5)$$

$$\min H(y^n(t), p^n(t), u(t)) = H^n(y(t), p(t), u(t)) \qquad (6)$$

Taylor's Theorem[5] admits the following representation of H(y(t), p(t), u(t), namely $H(y,p.u) = \sum_{i=0}^{k-1} \frac{1}{i!} <(y-y^n, p-p^n, u-u^n), \nabla>^i H^n(y,p.u) + \frac{1}{k!}$

$$<(y-y^n, p-p^n, u-u^n), \nabla>^k \; H^q(y,p,u) \qquad (7)$$

where $H^q(y,p.u) = H(y^q, p^q, u^q)$ and ∇ is the gradient operator; if H(y,p,u) and all its partial derivatives up through order k are continuous in some neighborhood N of (y^n, p^n, u^n). The vector (y^q, p^q, u^q) is a point on the line segment connecting $(y,p,u) \in N$. For obvious reasons the last term in (7) shall henceforth be referred to as the truncation error and denoted $o(\varepsilon^k)$.

The study shall now concentrate on a special truncated Taylor Series representation of the Hamiltonian; namely k=3. Therefore, (7) is being considered for k=3, and $H(y,p,u) \in C^3\ [t_o,T]$ at least, in some neighborhood of (y^n,p^n,u^n). Again let the cost functional be defined by (2) and the differential side constraint be given by

$$\dot{y} \triangleq g(y,u) = \left[f(y,u) \vdots 0 \right]^T, y_o \text{ given} \tag{8}$$

Notation: Let

$$\partial^k H^n(y,p,u) \Big/ \partial u^h \partial p^i \partial y^j \triangleq H^n_{y^j p^i u^h}$$

where $j + i + h = k$

The necessary conditions for the minimization of (2)subject to (8) become:

$$\dot{y} = \partial H(y,p,u)/\partial p = H^n_p + H^n_{p^2}(p-p^n) + H^n_{py}(y-y^n) + H^n_{pu}(u-u^n) + o_p(\varepsilon^3) \tag{9}$$

$$\dot{p} = -\partial H(y,p,u)/\partial y = H^n_y - H^n_{y^2}(y-y^n) - H^n_{yp}(p-p^n) - H^n_{yu}(u-u^n) - o_y(\varepsilon^3) \tag{10}$$

Let it be required, from $H(y,p,u) \in C^3\left[t_o,T\right]$, that the optimal control u^* satisfies

$$H_u = 0 \text{ for } u=u^* \in N \subset i(\Omega),\ \alpha \in N \tag{11}$$

where $i(\Omega)$ denotes the interior of Ω. Also let it be required that the second derivative type test $H_{u^2}(y,p,u) > 0$, thereby insuring a minimum. It should be noted that $H_u(y,p,u)$ and $H_{u^2}(y,p,u)$ exist over N. Therefore

$$H_u\Big|_{u=u^*} = H^n_u + H^n_{u^2}(u-u^n) + H^n_{uy}(y-y^n) + H^n_{up}(p-p^n) + o_u(\varepsilon^3)\Big|_{u=u^*} = 0 \tag{12}$$

The assumptions placed on u* are really not too restrictive if the class of control functions to be implemented possess some natural smoothness. However, such control strategies as bang-bang would naturally be excluded from this analysis.

From the extended canonical equations, several important properties pertaining to the extended Hamiltonian system become apparent. First, it is desirable to insure that the solutions to the extended canonical equations are well behaved for an arbitrary parameter vector α "close" to α^n.

Lemma 2-1 For $\alpha \to \alpha^n$, the solution of (9) converges uniquely to $y^n(t)$, the solution of (10) converges uniquely to $p^n(t)$ and the u which satisfies $H_u(y,p,u) = 0$ converges uniquely to $u^n(t)$.

Proof: By hypothesis there exists a unique control effort, call it $u_\alpha(t)$, which satisfies the original optimization problem with α being an arbitrary plant parameter and $\alpha \varepsilon \mathcal{A}$. Remembering (7) is an exact quality and for $(y,u) \varepsilon N$, $t \varepsilon \left[t_o,T\right]$, (i) $\dot{x}(t) = f(x(t),u_\alpha(t),\alpha)$, $x^n(t_o) = x^n_o$

Because for $\alpha = \alpha^n$, $f(x(t), \alpha^n, u^n(t))$ is Lipschitizian in x(t), and $u^n(t)$ (unique by hypothesis), the solution of (i), for $x(t_o) = x_o$, is unique and was defined to be $x^n(t)$. The function $f(x(t), \alpha, u_\alpha(t))$ is also continuous on Π_α,

$$\Pi_\alpha = \{ (t,x(t),\alpha) \mid t \varepsilon \left[t_o,T\right],\ (x,\alpha) \in N \}$$

Also, suppose $| f(x(t), \alpha, u_\alpha(t)) | \leq M$ on Π_α.

Claim: There exists a $\delta > 0$ such that for any fixed α (which implies a fixed $u_\alpha(t)$) with $||(\alpha - \alpha^n, u_\alpha - u^n)|| < \delta$, every solution ϕ_α of (i) exists uniquely on $[t_o, T]$ and as $\alpha \to \alpha^n$ (correspondingly $u_\alpha \to u^n$ from the continuity of $H_u(y,p,u)$), $\phi_\alpha(t) \to x^n(t)$.

Proof: The proof will be developed locally then extended. Choose λ sufficiently small so that

$$\Gamma \triangleq \{(t,x) \in E^{n+1} \mid [t-t_o] \leq \lambda \ , \ [x-x_o] \leq M\lambda, \text{ where } M \text{ is a Lipschitiz constant}\} \subset N.$$

Then, for any α, $u_\alpha \in N$ a solution of (i) exists on $[t-t_o] < \lambda$. Let ϕ_α be any solution of (i) (in the Caratheodory sense) for a parameter vector α and corresponding control u_α. Then $\{\phi_\alpha\}$ forms a bounded equicontinuous family. That is, for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$||\phi_\alpha(t) - \phi_\alpha(t')|| < \varepsilon$$

where $|t - t'| < \delta$. This implies, by Ascoli's Theorem[6], $\{\phi_\alpha(.)\}$ has a uniformly convergent subsequence $\{\phi_{\alpha(k)}(.)\}$

But from the uniqueness of the solution at $\alpha^n, u_\alpha^n = u^n$,

$$\phi_{\alpha(k)} \to \phi_\alpha n = x^n(t) \text{ as } \alpha(k) \to \alpha^n.$$

This gives a local result over $|t-t_o| < \lambda$. Extend the results to $[t_o, T]$. Notice also, only the uniqueness of ϕ_α at a point was required. #(Claim)

Thus, from the claim the solution of (i) tends to the unique solution $x^n(t)$ as $\alpha \to \alpha^n$. Obviously the solution of $\dot{\alpha}(t) = 0$, $\alpha(t_o) = \alpha$ tends to α^n as $\alpha \to \alpha^n$.

Now all that remains is to show $p(t) \to p^n(t)$ as $\alpha \to \alpha^n$. But from (ii) $p(t) = -H_y(y(t), p(t), u(t))$, $p(T)$ given and $H_y(y(t), p(t))$ is Lipschitzian in p. Therefore, there exists a unique solution of (ii) for $y(t)$, $u(t)$ assuming the role of the parameter vector α in the claim. Then, from the claim

$$p(t) \to p^n(t) \text{ as } \alpha \to \alpha^n \quad *$$

The previous results can be used to show how the cost index C(u) reacts as $\alpha \to \alpha^n$. The obvious question to ask is "does C(u) converge to C(u) as $\alpha \to \alpha^n$?"

Theorem 1: As $\tilde{\alpha} \to \alpha^n$, for $\tilde{\alpha}$, $\alpha^n \in \mathcal{A}$, $C(\tilde{u}) \to C(u^n)$, where $\tilde{u}$ denotes the unique solution of (12) for $\alpha = \tilde{\alpha}$.

Proof: All that need be proven is that $\tilde{x}(t) \to x^n(t)$ and $\tilde{u}(t) \to u^n(t)$. ($\tilde{x}$ is the solution of (2-9) for $\alpha = \tilde{\alpha}$)
Lemma 2-1 satisfies the demand. Thus $(\tilde{x}, \tilde{u}) \to (x^n, u^n)$ as $\tilde{\alpha} \to \alpha^n$, thus $C(\tilde{u}) \to C(u^n)$. *

What of the case where $\alpha \neq \alpha^n$? If the optimization problem is to make sense, one would intuitively expect the following to be trivially true.

Theorem 2, $\min C(u) = C(u^*) \leq C(\hat{u})$, where u* denotes the unique solution of (12) for $\alpha^* \neq \alpha^n$.

Proof:

Suppose not. Then there exists a $u^n \varepsilon N$ such that

$$C(u^n) < C(u^*).$$

But, for $(y^*, p^*, u^*) \varepsilon N$ the min $(H(y^*, p^*, u) = H(y^*, p^*, u^*), u \varepsilon i(\Omega)$. #

For the sake of notational convenience the following notational shall be established

Let $\Delta y = y - y^n$; $y, y^n \varepsilon N$

$\Delta p = p - p^n$; $p, p^n \varepsilon N$

$\Delta u = u - u^n$; $u, u^n \varepsilon N$.

Also, in (9), (10), or (12) it should be noted that $H_u^n(y,p,u) = \mathbf{0}$ and $H_u^n 2(y,p,u) = 0$ (due to p appearing linearly in $H(y, p, u)$). From (12), $\nabla^p u$ may be explicitly solved in terms of ∇y and ∇p because $H_u^n 2(y,p,u) > 0$ by hypothesis; therefore, $H_u^{n2}(y,p,u)^{-1}$ exists.

$$\Delta u = -\left[H_{u^2}^n\right]^{-1}\left[H_{py}^n \ \Delta y + H_{up}^n \ \Delta p + o_u(\varepsilon^3)\right]. \tag{13}$$

The results of (13) may now be used to eliminate Δu from (9) and (10) by direct substitution

$$\Delta\dot{y} = \left[H_{py}^n - H_{pu}^n\left[H_{u^2}^n\right]^{-1} H_{uy}^n\right] \Delta y - \left[H_{pu}^n\left[H_{u^2}^n\right]^{-1} H_{up}^n\right]\Delta p +$$

$$-\left[H_{pu}^n\left[H_{u^2}^n\right]^{-1}\right]o_u(\varepsilon^3) + o_p(\varepsilon^3) \tag{14}$$

where: $\Delta y(t_o) = \left[\frac{0}{\gamma}\right]$,

γ arbitrary.

The arbitrariness of γ will not affect the desired result as will be shown shortly and

$$\Delta\dot{p} = \left[H_{y^2}^n - H_{yu}^n\left[H_{u^2}^{n\ -1}\right] H_{uy}^n\right] \Delta y - \left[H_{yu}^n \ H_{u^2}^{n\ -1} \ H_{up}^n\right] \Delta p +$$

$$-\left[H_{yu}^n\left[H_{u^2}^{n\ -1}\right]\right]o_u(\varepsilon^3) - o_y(\varepsilon^3), \tag{15}$$

$\nabla p(T)$ given.

For N sufficiently small such that the higher order terms $o_u(\varepsilon^3)$, $o_p(\varepsilon^3)$ are negligible, then (14) and (15) can be represented as a system of $2_3(n+m)$ linear ordinary differential equations. (Error bounds on $o_u(\varepsilon^3)$, $o_p(\varepsilon^3)$ and $o_y(\varepsilon^3)$ will be established later). Suppose such a non-empty neighborhood exists.

Then:

$$\begin{bmatrix} \Delta\dot{y} \\ \Delta\dot{p} \end{bmatrix} = \left[\begin{array}{c|c} A(t) & B(t) \\ \hline C(t) & -A^T(t) \end{array}\right] \begin{bmatrix} \Delta y \\ \Delta p \end{bmatrix} \triangleq G(t) \begin{bmatrix} \Delta y \\ \Delta p \end{bmatrix} \tag{16}$$

Δy_0 $\quad \Delta p(T)$ given

where A, B, and C are defined in the obvious manner.

For $H(y,p,u) = L(x,u) + \langle \bar{p}, f(x,u,\alpha) \rangle + \langle p, 0 \rangle$

where $p = \begin{bmatrix} \bar{p}\ (nx1) \\ \hline p\ (mx1) \end{bmatrix}$, $p(T) = \begin{bmatrix} \text{given} \\ \hline 0 \end{bmatrix}$

A rather interesting phenomenon occurs in equation (16). The rows of its $\Delta\dot{\alpha}$ partition are equal to zero. In fact, the truncation errors associated with the $\Delta\dot{\alpha}$ rows are also zero. This implies that the autonomy of the parameters have been analytically preserved. Therefore, the parameter variations considered, of the form $\alpha - \alpha^n$, equal a constant.

Eq. (16) is recognized to be a matrix Riccati type equation. It can be recognized as such because of the structure of G(t). The solution of (16) is given by a (n+m) x (n+m) nonlinear system of differential equations. That is, there exists a symmetric K(t) such that

$$-\dot{K}(t) = A^T(t)\,K(t) + K(t)\,A(t) - K(t)\,B(t)\,K(t) + C(t)$$

$$K(t) = \begin{bmatrix} \Gamma & 0 \\ 0 & 0 \end{bmatrix}$$

and $\Delta p(t) = K(t)\,\Delta y(t)$. (17)

The solution of (17) will usually involve a computer mechanization of the problem. Some techniques of solving a Riccati equation can be found in Friedland [7], and Bass with an example supplied by Stoner, Taylor, and Bass [8]. If any of the 2(n+m) solutions of (16) are known in closed form, then partioning routines can be used to reduce the computational difficulties associated with finding K(t) [9]. Partioning techniques also suggest another method of realizing the elements of K(t) for this particular problem.

Partition K(t) follows:

$$K(t) = \left[\begin{array}{c|c} K_1(t) & K_2(t) \\ \hline K_2^T(t) & K_3(t) \end{array}\right]$$

where K_1 is nxn, K_2 is nxm, K_3 is mxn. Then ∇u defined by (13 & 17) can be expressed as $L(\nabla y)$, where L is a linear operator. Therefore, one needs only to focus attention on calculating $K_1(t)$ and $K_2(t)$.

From calculating A, B, C from (16), the following results are achieved

$$A(t)= \begin{matrix} n \\ m \end{matrix} \left\{ \begin{bmatrix} \underbrace{A_1(t)}_{n} & \underbrace{A_2(t)}_{m} \\ \underbrace{0}_{n} & \underbrace{0}_{m} \end{bmatrix} \right., \quad B(t) = \begin{matrix} n \\ m \end{matrix} \left\{ \begin{bmatrix} \underbrace{B_1(t)}_{n} & \underbrace{0}_{m} \\ \underbrace{0}_{n} & \underbrace{0}_{m} \end{bmatrix} \right.,$$

$$C(t) = \begin{matrix} n \\ m \end{matrix} \left\{ \begin{bmatrix} \underbrace{C_1(t)}_{n} & \underbrace{C_2(t)}_{m} \\ \underbrace{C_3(t)}_{n} & \underbrace{C_4(t)}_{m} \end{bmatrix} \right.$$

Therefore:

$$\text{(i)} \quad -\dot{K}_1(t) = K_1(t)\, A_1(t) + A_1^T(t)\, K_1(t) - K_1(t)\, B_1(t) K_1(t) + C_1(t), \quad K_1(T) = \Gamma$$

and

$$\text{(ii)} \quad -\dot{K}_2(t) = K_1(t) A_2(t) + A_1^T(t) - K_1(t)\, K_2(t) + C_2(t), \quad K_2(T) = 0.$$

Comments:

Equation (i) is a homogeneous nxm nonlinear differential system of equations. Its boundary data is specified at t = T. It is independent of the initial data $\nabla y(t_0)$.

Equation (ii) is a nonhomogeneous (nxm) linear differential system of equations. Its boundary data is specified at t=T. It is independent of the initial data $\nabla y(t_0)$, thus leaving $\nabla\alpha_0$ arbitrary is acceptable.

Consider now the solution of (16) to be of the form

$$\nabla u(t) = \tilde{G}\, \nabla y(t) \tag{19}$$

where $\tilde{G}(t)$ is calculated through one of the suggested methods. The implementation of the adaptive control need only require that $x^n(t)$ need be stored in some memory device. The simplest example of such a device would be a tape. Thus, after $\tilde{G}(t)$ has been precomputed, $\tilde{G}(t)$, being a function of the nominal variables $y^n(t)$, $p^n(t)$ and $u^n(t)$, it will be multiplied in real time by $(y(t)-y^n(t))$. But $y^n(t)$ equals:

$(x^n(t), \alpha^n)$ where $x^n(t)$ will be stored in the memory device and α^n is nothing more than some known constant biasing term. Thus only n known time varying functions need be introduced into the system in real time instead of 2(m+m)+r function which might have been earlier predicted

The schematized realization of the adaptive system is given in Figure 1, where α^e is an estimation of α^* and α^* is the actual plant parameter vector.

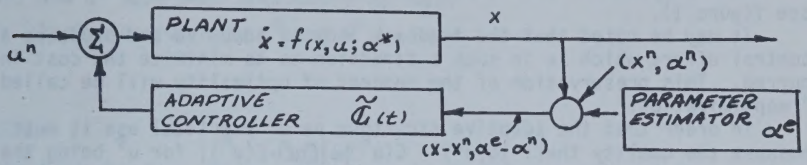


Figure 1
Adaptive Structure

Truncation Errors: Rather than an exhibit the results in tensor notation, the equivalent scalar triple sum will be analyzed then the truncation error of (7) becomes:

$$(i)\quad \varepsilon = \frac{1}{3!} \sum_{i,j,k=1}^{2(n+m)+r} (z_i - z_i^n)(z_j - z_j^n)(z_k - z_k^n) \frac{\partial^3 H^q(z)}{\partial z_i \partial z_i \partial z_k}, \quad z \triangleq (y,p,u)$$

The requirement of finding a z^q which establishes the equality of (7) may be relaxed by defining ε' to be the truncation error.

$$(ii)\quad \varepsilon \le \varepsilon' = \sup_{z \in N} \frac{1}{3!} \sum_{i,j,k=1}^{2(n+m)+r} (z_k - z_k^n)(z_i - z_i^n)(z_j - z_j^n) \frac{\partial^3 H(z)}{\partial z_i \partial z_j \partial z_k}$$

Also, the truncation error associated with (13), (14) and (15) may be interpreted similarly.

Then
$$(iii)\quad o_u(\varepsilon^3) \le \sup_{z \in N} \left| \frac{\partial \varepsilon'}{\partial u} \right|, \; o_p(\varepsilon^3) \le \sup_{z \in N} \left| \frac{\partial \varepsilon'}{\partial p} \right|, \; o_y(\varepsilon^3) \le \sup_{z \in N} \left| \frac{\partial \varepsilon'}{\partial y} \right|$$

Therefore, by requiring $o_u(\varepsilon^3)$, $o_p(\varepsilon^3)$ and $o_y(\varepsilon^3)$ being small can be insured by requiring the right hand side of (iii) be small.

EXAMPLE PROBLEM:
Consider the plant $\dot{X} = \alpha x(t) + u(t), x(0) = 1, \alpha^n = -1$, and $C(u) = \frac{1}{2}\int_o^\infty (X^2 + U^2)dt$. Performing the required computation on finds:

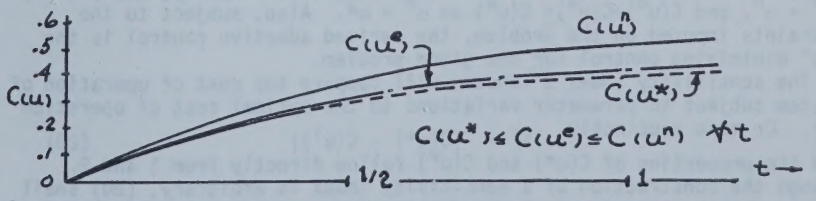
For $\alpha = \alpha^n - \varepsilon$, $A \triangleq (1 - \sqrt{2})$, $B \triangleq (1 + \varepsilon)$,

$u = u^n = Ae^{(-\sqrt{2}t)} \rightarrow X = \exp[-(1+\varepsilon)t][1 + (A/(A+\varepsilon))(\exp(A+\varepsilon)t - 1)]$

$u = u^e = u^n - [-A(x-x^n) - \exp(-\sqrt{2}t)\, 2A\varepsilon] \rightarrow x^e = \exp[-2+\varepsilon)t]\{1 + 2A(\exp(\varepsilon t) - 1\}$.

$u = u^* = -B + \sqrt{B^2+1}\, x^* \rightarrow x^* = \exp(-\sqrt{B^2+1}\, t)$

For $\varepsilon = o$ the variables x and u degenerate into x^n and u^n. For $\varepsilon = .1$ (ie. 10% parameter variation) the results exhibited in figure 2 are achieved.



Reoptimization It is assumed: (i) that the actual state $x(t)$ can be measured in real time in the absence of noise. (ii) $x^n(t)$ and α^n can be loaded into the system in real time, and (iii) α^n can be generated (estimated, see figure 1).

It may be noted that the feedback loop is adaptive and produces a control effort which is in such a direction as to minimize the cost incurred. This preservation of the concept of optimality will be called "reoptimization."

In order that the adaptive structure be of practical use it must possess the quality that: $|C(u^*) - C(u^e)| \leq |C(u^*)-C(u^n)|$ for u^e being the adaptive control and for a set of parameter vectors in some neighborhood of α^n.

Theorem 3: For an adaptive control u^e satisfying (19) $u^e \varepsilon i(\Omega)$, there exists $\varepsilon \geq 0$, and sufficiently small, such that for $||\alpha^e - \alpha^*|| < \varepsilon; \alpha^e, \alpha^* \varepsilon \alpha$,

$$|C(u^*) - C(u^e)| \leq |C(u^*) - C(u^n)|.$$

Proof: For $\alpha = \alpha^* \neq \alpha^n, C(u^*) < C(u^n)$ from Theorem 2. From (14) and the matricies defined by (16) (α^e being the approximation of α^*) (*) $\dot{y}^* - \dot{y}^e = A(y^* - y^e) + B(p^* - p^e) +$
$+ \gamma (y^* - y^n, p^* - p^* - p^n, u^* - u^n), (y_0^* - y_0^e) = \left|\begin{matrix}\rho \\ \sigma\end{matrix}\right|$

where $\gamma (y^* - y^n\ p^* - p^n, u^* - u^n)$ is the truncation error of (14). From (15) and the matrices defined by (16) $\dot{\Phi}(p^*-p^e) = C(y^* - y^e) - A^t(p^*-p^e) +$
$+ \bar{\gamma}\ (y^* - y^n, p^* - p^n, u^* - u^n), (p^*(T) - p^e(T)) =$ given

Now choose $N_\varepsilon \neq \phi$ such that the truncation errors γ and $\bar{\gamma}$ are neglectable. Either N_ε is the point set (y^n, p^n, u^n) or otherwise. If N_ε is the point set (which is the result of insufficient smoothness of $H(y,p,u)$ about $H^n(y,p,u)$, finished. If N_ε is not the point set, then there exists α^e such that $||\alpha^* - \alpha^e|| < ||\alpha^* - \alpha^n||$. In an argument similar to that of Lemma 2-1, the solution of

$$(i)\begin{bmatrix} \dot{y}^* - \dot{y}^e \\ \dot{p}^* - \dot{p}^e \end{bmatrix} = \left[\begin{array}{c|c} A(t) & B(t) \\ \hline C(t) & -A^t(t) \end{array}\right] \begin{bmatrix} y^* - y^e \\ p^* - p^e \end{bmatrix}$$

tends to the trivial solution $(y^*(t) - y^e(t)) \to 0$ and $(p^*(t) - p^e(t)) \to 0$. The trivial solution is the unique solution of (i) for $\alpha^e = \alpha^*$ (i.e., $\varepsilon = 0$). It is apparent that u^e tends to u^* for this local argument because of their linear structure.

Therefore, for α^* and $\alpha^e \varepsilon N_\varepsilon$ and α^e sufficiently close to α^*, $L(x^e, u^e)$ defined in (2) can be made arbitrarily close or equal to $L(x^*, u^*)$. But u^n is a fixed function of time and for $\alpha^* \neq \alpha^n, \alpha^* - \alpha^n$ is some constant non zero vector λ.

Consider the sequency of parameter vectors $\alpha_i^n \varepsilon N_\varepsilon$, such that $||\alpha_i^n - \alpha_n^n|| = 0$ for $i = 1, 2 \ldots$. Then for α^* and $\alpha_i^n \varepsilon N_\varepsilon$, and $||\alpha^* \varepsilon - \alpha_i^n|| = \lambda$, $L(x_i^n, u_i^n)$ need not converge to $L(x^*, u^*)$ as $i \to \infty$. Then $|C(u^*) - C(u^e)| \leq |C(u^*) - C(u^n)|$ for a set of α^e's sufficiently to α^*.

The adaptive structure now formally processed the qualities demanded of a reoptimization scheme (see Theorem (1), (2), (3)). Namely, $C(u^e) \to C(u^n)$ as $\alpha^e \to \alpha^n$, and $C(u^e) \to C(u^*) \leq C(u^n)$ as $\alpha^e \to \alpha^*$. Also, subject to the constraints imposed on the problem, the derived adaptive control is the "best" minimizing control for the given problem.

The sensitivity index S^γ chosen will compare the cost of operation of a system subject to parameter variations to the optimal cost of operation $C(u^*)$. Or more succinctly

$$S^\gamma = |C(u^*) - C(u^\gamma)| \qquad (20)$$

where the properties of $C(u^*)$ and $C(u^\gamma)$ follow directly from 1 and 2. Although the construction of a sensitivity index is arbitrary, (20) shall be assumed to be the canonical sensitivity index. With this definition

of S^{γ} and Theorem (3) in mind, the following is easily shown.

Correlary (3): Under the conditions state in Theorem 3, $S^{e} \leq S^{n}$ where: u^{*}, u^{e}, and u^{n} have been previously defined.

Proof: Apply Theorem 3 to S^{e} and S^{n}.

Summary:

As a synthesis technique, this result could be used to design an efficient control systems for certain engineering plants or industrial processes whose parameter are subject to large variations during their operation. Finally, the true structure of the adaptive optimal controller in relation to parameter changes, which had not previously been derived, has a simple structure whose only drawback is that the plant parameters must be estimated. This, however, is a separate problem area which has, and is currently, receiving attention.

References:

1. Kelly, H. J., "Guidance Theory and Extremal Fields," IRE Trans on Auto Control, vol, AC7, Oct. 1962 pp 75-82.
2. Breakwell, Bryson, and Speyer, "Optimization and Control of Nonlinear System Using the Second Variation," Journal S.I.A.M. on Control, Series A, vol 1, no 2, 1963 pp 193-223.
3. Friedland, B, "A Technique of Quasi-Optimum Control," Transactions of the A.S.M.E. Basic Engineering,June 1966, pp 437-443.
4. Pontryagin, Boltyanski, Jamkrelidze, Mischenko, Mathematical Theory of Optimal Control Processes, Intersience Pub., p75,1962.
5. Fulks, W. Advanced Calculus, Wiley, p 230. 1961.
6. Simmons, G., Topology and Modern Analysis,-McGraw Hill, pp 126-128,1963.
7. Friedland, B., "On Solutions of the Ricatti Equation in Optimization Problems, "IEEE Trans on Automatic Control, Vol AC-1 pp 303-304, June 1967.
8. Bass, R. W. "Machine Solution of High Order Matrix Riccati Equations, Douglas Missle and Space Division Pub., 1967.
9. Haritman, P., Ordinary Differential Equations, John Wiley and Sons, Chapter V, 1964.

AN ADAPTIVE APPROACH TO THE TRAJECTORY SENSITIVITY DESIGN OF OPTIMAL REGULATORS *

SIRA G. RAO and A.C. SOUDACK **
Electrical Engineering Department
University of British Columbia, Vancouver, B.C.

ABSTRACT

Often in deterministic linear optimal regulator problems, the plant parameters may not be exactly known or they might have changed from their nominal values. An effective method of design in such situations is to minimize a cost term of trajectory sensitivity along with the originally selected performance index. A design procedure is proposed here for this purpose.

COMPUTATION OF TRAJECTORY SENSITIVITY

The plant state equation is given by

$$\dot{\underline{x}} = A\underline{x} + B\underline{u}; \quad \underline{x}(t_o) = \underline{x}_o \tag{1}$$

In eqn. (1), $\underline{x}$, the state vector, is n-dimensional, $\underline{u}$ is an m-vector, A and B are $(n\times n)$ and $(n\times m)$ matrices, and $\underline{x}_o$ is the initial value of $\underline{x}$. Let α be a constant parameter whose exact value is unknown. (The procedure given below can be generalized to the case where more than one uncertain parameter is present.) The trajectory sensitivity $\underline{q}$ is defined by the partial derivative of $\underline{x}$ with respect to α. That is,

$$\underline{q} \triangleq (\partial \underline{x}/\partial \alpha) \tag{2}$$

The optimal control problem is to select a $\underline{u}^*$, if it exists, such that a performance index J, defined by

$$J = \int_{t_o}^{t_f} (\underline{x}^T Q_1 \underline{x} + \underline{q}^T Q_2 \underline{q} + \underline{u}^T R \underline{u}) dt; \; Q_1, \; Q_2 \geq 0; \; R > 0; \tag{3}$$

is minimized for $\underline{u} = \underline{u}^*$. In the general case $\underline{u}^*$ is a vector function $\underline{x}$, $\underline{q}$, t and t_o:

$$\underline{u}^* = \underline{h}\,(\underline{x}, \; \underline{q}, \; t, \; t_o) \tag{4}$$

One attempts to obtain $\underline{q}$, which is required along with measured values of $\underline{x}$, t and t_o, in the implementation of the control law of eqn. (4), by taking partial derivatives of eqn. (1) with respect to α. The result is,

$$\dot{\underline{q}} = A\underline{q} + A_\alpha \underline{x} + B_\alpha \underline{u}^* + B\underline{u}_\alpha^*; \; \underline{u}_\alpha^* \triangleq (\partial \underline{u}^*/\partial\alpha); \; A_\alpha \triangleq (\partial A/\partial \alpha) \tag{5}$$

From equation (4), upon taking partial derivatives,

---

*This research was supported by the National Research Council of Canada Grant No. 67-3138.

**Present Address: Visiting Associate Professor, Technion, Haifa, Israel.

$$\underline{u}_\alpha^* \doteq (\underline{h}_{\underline{x}})^T \underline{q} + (\underline{h}_{\underline{q}})^T \underline{q}_\alpha; \quad \underline{q}_\alpha \triangleq (\partial \underline{q}/\partial \alpha) \tag{6}$$

A new unknown, $\underline{q}_\alpha$, is present in eqn. (6), which one may attempt to obtain, by taking partials of eqn. (5) w.r.t.α. In the process, yet another unknown $\underline{q}_{\alpha\alpha}$ is introduced, which cannot be neglected. To bypass such difficulties of computation, a successive approximation procedure is proposed:

SUBOPTIMAL CONTROL

Rewriting eqn. (5) in the form

$$\dot{\underline{q}} = A\underline{q} + A_\alpha \underline{x} + B_\alpha \underline{u} + B\underline{z}; \quad \underline{z} \epsilon R^m \tag{7}$$

the optimization problem may be posed as the requirement to minimize J subject to the constraint that $\underline{u}_\alpha = \underline{z}$, a.e. In this case, the optimal solution cannot be determined, however, by known techniques. Instead one considers the suboptimal solution $\underline{u}^*$, which is optimal for a performance index J_1, where

$$J_1 = \int_{t_o}^{t_f} (\underline{x}^T Q_1 \underline{x} + \underline{y}^T Q_2 \underline{y} + \underline{u}^T R \underline{u})\, dt \tag{8}$$

subject to the equation

$$\dot{\underline{y}} = C\underline{y} + D\underline{x}; \quad \underline{y}(t_o) = \underline{0}; \quad \underline{y} \epsilon R^n \tag{9}$$

and simultaneously minimizes

$$J_2 = \int_{t_o}^{t_f} [(\underline{y} - \underline{q})^T (\underline{y} - \underline{q})]\, dt \tag{10}$$

If now one assumes a suboptimal solution of the form

$$\underline{u}^* = - R^{-1} B_1^T K \underline{s}; \quad \underline{s}^T \triangleq (\underline{x}^T \underline{y}^T); \quad B_1^T = (B^T \; 0) \tag{11}$$

where K is the solution of the Ricati equation corresponding to the system equation (1) and (9) and the performance index J, a gradient technique may be applied to the adjustment of C and D such that J_2 is small.

Thus, if c_i refers to a parameter of the system of equation (9), then the increment in c_i is given by, for $\lambda_i > 0$,

$$\Delta c_i = -\lambda_i \frac{\partial J_2}{\partial c_i} = -\lambda_i \int_{t_o}^{t_f} [(\underline{y}_i - \underline{q}_i)^T (\underline{y}_i - \underline{q}_i)]dt; \; \underline{y}_i \triangleq (\partial \underline{y}/\partial c_i); \; \underline{q}_i \triangleq (\partial \underline{q}/\partial c_i); \tag{12}$$

The quantities $\underline{q}$, $\underline{q}_i$, $\underline{y}_i$ are obtained from equations (1), (9), (11) and the Ricati equation by taking appropriate partial derivatives. With Δc_i changed thus, the procedure is repeated for the new value of c_i and continued until J_2 is small. The above procedure can be easily extended to the design of time optimal control systems.

REFERENCES

1. S.G. Rao and A.C. Soudack, "Synthesis of Optimal Control Systems with Near Sensitivity Feedback", (submitted for publication).
2. E. Kreindler, "Optimization with Sensitivity Considerations", Grumann Research Report RE-305, October 1967, Grumman Aircraft Corporation, New York.

SOME RESULTS ON THE USE OF REDUCED MODELS IN THE CONTROL OF LINEAR DYNAMICAL SYSTEMS

D. MITRA
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

ABSTRACT

Two sets of first order differential equations denoted by X and X_r are assumed to exist where X and X_r respectively govern the behavior of the plant and the reduced model. The investigation is on the performance of the plant when it is controlled on the false assumption that its behavior is governed by X_r.

Three specific optimal control problems are considered and in each case explicit relations are obtained for the deterioration in performance in model control of the plant. In the first two problems the interconnection between the error functional commonly adopted for the model reduction problem and the deterioration in performance is demonstrated; the results are applied to the special case of reduced models obtained by the method of optimal projection along invariant subspaces.

1. INTRODUCTION

This paper summarizes the results reported in [10]. Two sets of first order differential equations denoted by X and X_r are assumed to exist where X and X_r respectively govern the behavior of the plant and the reduced model. The investigation is on the suboptimal performance of the plant when it is controlled on the false assumption that its behavior is governed by X_r. Three optimal control problems are considered; in each case the control function derived on the basis of the reduced model equations is applied to the plant and explicit expressions are obtained for the resulting deterioration in performance.

A method for reducing the order of linear, time-invariant dynamical models has been obtained by the author. Details of this method of reduction has been reported elsewhere [1]. The application of general results to the particular case of reduced models obtained by this method is considered. Numerical results on the reduction of a ninth order linearized model of a boiler is reported in [2]. Further numerical results on the topics considered here are contained in [3].

1.1 Preliminaries

The two sets of first order linear, time-invariant differential equations denoted by X and X_r are -

$$X : \frac{d}{dt}x = Ax + Bu(t) \qquad (1,a)$$

$$y(t) = Cx(t)$$

$$X_r : \frac{d}{dt}x_r = A_r x_r + B_r u(t) \qquad (2)$$

x, x_r, y and u are assumed to be n, n-m, n-m and p dimensional vectors with $m \geq 0$. X and X_r are assumed to be completely controllable. C is assumed to be of rank (n-m). X_r is such that $x_t(t)$ is an approximation of y(t), the output of the system X. It should be noted that the form of X_r is a special case of the general form of reduced systems usually considered [1].

Certain matrices are best defined at the outset. H is a p by p symmetric, positive definite matrix.*

$$R(t_1,t) = e^{A(t_1-t)}B \tag{3}$$

$$R_y(t_1,t) = Ce^{A(t_1-t)}B \tag{4}$$

$$R_r(t_1,t) = e^{A_r(t_1-t)}B_r \tag{5}$$

$$W(t_1,t_0) = \int_{t_0}^{t_1} R(t_1,t)H^{-1}R^t(t_1,t)dt \tag{6}$$

$$W_y(t_1,t_0) = \int_{t_0}^{t_1} R_y(t_1,t)H^{-1}R_y^t(t_1,t)dt = CW(t_1,t_0)C^t \tag{7}$$

$$W_c(t_1,t_0) = \int_{t_0}^{t_1} R_y(t_1,t)H^{-1}R_r^t(t_1,t)dt \tag{8}$$

$$W_r(t_1,t_0) = \int_{t_0}^{t_1} R_r(t_1,t)H^{-1}R_r^t(t_1,t)dt \tag{9}$$

For $t_0 = 0$ the set of matrices $W(t_1,t_0)$ will sometimes be denoted by $W(t_1)$.

The set of W matrices have been discussed elsewhere [4]. It has been shown that W and W_y are the Gramian matrices associated respectively with the state and output vectors of the system X; likewise W_r is the Gramian matrix for the state vector of X_r. W_c is the coupling matrix linking the output vector of X with the state vector of X_r. The matrices W and W_r satisfy matrix Riccati equations. Complete controllability of X and X_r implies that W and W_r are positive definite. Since the rank of the (n-m) by n matrix C is (n-m), W_y is also positive definite. W_c in general is neither positive definite nor symmetric.

1.2 Reduction by Optimal Projection Along an Invariant Subspace

It may be shown [10] that the following results hold when the reduced model X_r is obtained from X by the method of optimal projection,

$$W_r(t_1) = W_y(t_1) - (CT_1)[T_1^tW^{-1}(t_1)T_1]^{-1}(CT_1)^t \tag{10}$$

where the m columns of the matrix T_1 form an orthonormal set and span the subspace along which the projection takes place.

Also,
$$W_c(t_1) = W_r(t_1) \tag{11}$$

Equation (11) represents a fundamental property of this reduction procedure and the following inference may be drawn. From (8) and (9),

$$W_r(t_1) - W_c^t(t_1) = \int_0^{t_1} R_r(t)H^{-1}[R_r(t) - R_y(t)]^t dt \tag{12}$$

---

* The superscript t denotes the transpose of a matrix.

Hence (11) implies that

$$\int_0^{t_1} R_r(t)H^{-1}[R_r(t) - R_y(t)]^t dt = 0 \tag{13}$$

The above relation is the condition for an orthogonal projection in function space [5].

Consider W_r and W_c relative to W_y. It may be shown that $CT_1(T_1^t W^{-1} T_1^t)^{-1} T_1^t C^t$ is positive definite. It follows from (10) that

$$W_y > W_r = W_c > 0 \tag{14}$$

It is well known [6] that the above set of inequalities implies

$$W_c^{-1} = W_r^{-1} > W_y^{-1} \tag{15}$$

1.3 The Functional for the Reduction Process

One of the functionals to be minimized in the problem of model reduction as formulated in [1] is

$$E(t_1) = \sum_{i=1}^{p} \int_0^{t_1} \|r_y^i(t) - r_r^i(t)\|_Q^2 dt \tag{16}$$

where $r_y^i(t)$ is the response of the output of the system X to a weighted impulse applied at the i^{th} input. $r_r^i(t)$ denotes the state variable response of the system X_r to an identical input. The input weighting is determined by a positive definite matrix H^{-1}, say. Rearranging the right hand side expression in (16) yields

$$E(t_1) = \text{Trace } Q'[W_y(t_1) - 2W_c(t_1) + W_r(t_1)] \tag{17}$$

For $W_c(t_1) = W_r(t_1)$, (17) reduces to

$$E(t_1) = \text{Trace } Q'[W_y(t_1) - W_r(t_1)] \tag{18}$$

2. REDUCED MODELS IN SOME OPTIMAL CONTROL PROBLEMS

Three specific optimal control problems are considered. For each of these problems analytic expressions are obtained for three functionals J_1^0, J_2, and J_3^0, where the superscript 0 is a nmemonic for "optimal" and the numerical subscript refers to one of three configurations. The first configuration is one where the plant, which is assumed to be exactly governed by the set of differential equations denoted by X in (1), is subject to optimal control $u_1^0(t)$ which is derived for X. The third configuration is one where the model X_r is subject to optimal control $u_3^0(t)$ which is derived for X_r. Thus the first and third configurations are "ideal" systems. In the second configuration the plant is subject to $u_3^0(t)$. Essentially the plant is controlled in this configuration as though it were the model. There exists trajectories 1, 2, 3 associated with the appropriate configurations.

Obviously the main interest is in $(J_2 - J_1^0)$ which provides a measure of the sub-optimality introduced in control of this nature.

2.1 Fixed Target, Fixed Terminal Point, Minimum Energy Problem

(a) For the first configuration the functional is

$$J_1 = \int_0^{t_1} \|u_1(t)\|_H^2 dt \tag{19}$$

and it is required that

$$y(t_1) = y', \text{ given.} \tag{20}$$

It is given that $x(0) = 0$. Also H is assumed to be positive-definite. y' is the fixed target and complete controllability implies that some control exists which accomplishes the desired transfer.

The functional may be minimized to yield the optimal values of $u_1(t)$ which is denoted by $u_1^0(t)$ and its substitution in (19) yields the minimum functional J_1^0.

Kalman et al. [7] have studied this problem for the case when the target is in state-space rather than in output-space i.e. it is required that

$$x(t_1) = x', \text{ given.} \tag{21}$$

It has been shown [10] that their results may be modified in a straightforward manner. In fact,

$$u_1^0(t) = H^{-1} R_y^t(t_1,t) W_y^{-1}(t_1) y' \tag{22}$$

and

$$J_1^0 = \int_0^{t_1} \|u_1^0(t)\|_H^2 dt = \|y'\|^2_{W_y^{-1}(t_1)} \tag{23}$$

(b) For the third configuration it is reasonable to consider the functional

$$J_3 = \int_0^{t_1} \|u_3(t)\|_H^2 dt \tag{24}$$

with the requirement that

$$x_r(t_1) = y' \tag{25}$$

It is assumed that $x_r(0) = 0$.

The results of [7] are directly applicable; hence

$$u_3^0(t) = H^{-1} R_r^t(t_1,t) W_r^{-1}(t_1) y' \tag{26}$$

and

$$J_3^0 = \|y'\|^2_{W_r^{-1}(t_1)} \tag{27}$$

(c) Now consider the second configuration. The state and output of X at time t when X is subject to $u_3^0(t)$ is denoted by $x_2(t)$ and $y_2(t)$ respectively. In particular

$$x_2(t_1) = \int_0^{t_1} R(t_1,t) H^{-1} R_r^t(t_1,t) dt \; W_r^{-1}(t_1) y' \tag{28}$$

and $$y_2(t_1) = \int_0^{t_1} R_y(t_1,t)H^{-1}R_r^t(t_1,t)dt\, W_r^{-1}(t_1)y' = W_c(t_1)W_r^{-1}(t_1)y' \qquad (29)$$

(d) Now if e_{12} is defined to be the target error for trajectory 2, or the vector by which the plant output misses the target when it is controlled as though it were the reduced model,

$$e_{12} = y' - y_2(t_1) = \left[I_{n-m} - W_c(t_1)W_r^{-1}(t_1)\right] y' \qquad (30)$$

Equation (30) implies that $e_{12} = 0$, for arbitrary y', if and only if

$$W_c(t_1) = W_r(t_1). \qquad (31)$$

Also $$\int_0^{t_1} \left[\|u_3^0(t)\|_H^2 - \|u_1^0(t)\|_H^2\right]dt = y'^t \left[W_r^{-1}(t_1) - W_y^{-1}(t)\right] y' \qquad (32)$$

and,

$$\int_0^{t_1} \|u_3^0(t) - u_1^0(t)\|_H^2 dt = y'^t \left[W_y^{-1}(t_1) - 2W_y^{-1}(t_1)W_c(t_1)W_r^{-1}(t_1) + W_r^{-1}(t_1)\right] y' \qquad (33)$$

(e) A neat comparison of the two types of plant control represented by configurations 1 and 2 is possible by assuming the target is distributed over the output space with density

$$p(y') = \frac{1}{\sqrt{(2\pi)^{n-m}\det Q}} \exp -\tfrac{1}{2}(y'^t Q^{-1} y') \qquad (34)$$

where Q is a symmetric positive-definite matrix. Note that (34) implies*

$$M[y' <> y'^]\ = Q, \quad M[y'] = 0 \qquad (35)$$

From (32)

$$M\left[\int_0^{t_1} \{\|u_3^0(t)\|_H^2 - \|u_1^0(t)\|_H^2\} dt\right] = \text{trace } W_y^{-1}(W_y - W_r)W_r^{-1}Q \qquad (36)$$

Proceeding in a similar manner,

$$M\left[\int_0^{t_1} \|u_3^0(t) - u_1^0(t)\|_H^2 dt\right] = \text{Trace } W_y^{-1}(W_y - 2W_c + W_r)W_r^{-1}Q \qquad (37)$$

It has been shown that comparison between the two types of plant control is only meaningful if

$$W_c(t_1) = W_r(t_1) \qquad (38)$$

in which case, from (36) and (37),

$$M\left[\int_0^{t_1} \|u_3^0(t)\|_H^2 - \|u_1^0(t)\|_H^2 dt\right] = M\left[\int_0^{t_1} \|u_3^0(t) - u_1^0(t)\|_H^2 dt\right] = \text{trace } W_y^{-1}(W_y - W_r)W_r^{-1}Q \qquad (39)$$

---

* M is the mathematical expectation operator.

We now observe that it is reasonable to assume that Q is equal to either W_r or W_y. A physical interpretation of this assumption is that the probability of a linear combination of output variables being the target is such that its variance is directly proportional to its controllability. Equating Q to W_r in (39) we obtain for the mean of the control energy difference,

$$\text{trace } W_y^{-1}(W_y - W_r) \tag{40}$$

Furthermore, if

$$Q \doteq W_r W_y \tag{41}$$

substitution in (39) yields the following expression for the mean of the control energy differences -

$$\text{trace } W_y^{-1}(W_y - W_r)W_y = \text{trace}(W_y - W_r) \tag{42}$$

Equations (40) and (42) should be compared with the expression for the error functional in the reduction process with the constraint $W_c = W_r$, which is given in Eq. (18).

2.2 Fixed Target, Free Terminal Point Minimum Energy Tracking Problem*

(a) For the first configuration the functional is

$$J_1 = \|y' - y_1(t_1)\|_F^2 + \int_0^{t_1} \|u_1(t)\|_H^2 dt \tag{43}$$

where y' is a given target in the output space of X. F and H are respectively (n-m) positive semi-definite and p by p positive definite matrices. It is assumed that x(0) = 0.

It may be shown by a straightforward application of Pontryagin's Maximum Principle that [10]

$$u_1^0(t) = H^{-1}B^t A^t(t_1 - t)C^t F[y' - Cx_1(t_1)] \tag{44}$$

and the "target-miss",

$$y' - Cx_1(t_1) = [I_{n-m} + W_y(t_1)F]^{-1}y' \tag{45}$$

Thus

$$u_1^0(t) = H^{-1}R_y^t(t_1, t)F[I_{n-m} + W_y(t_1)F]^{-1}y' \tag{46}$$

A matrix V_y is defined thus,

$$V_y = F[I_{n-m} + W_y(t_1)F]^{-1} \tag{47}$$

Then

$$u_1^0(t) = H^{-1}R_y^t(t_1, t)V_y y' \tag{48}$$

The control energy is

$$\int_0^{t_1} \|u_1^0(t)\|_H^2 dt = y'^t V_y^t W_y(t_1) V_y y' \tag{49}$$

---

* Results similar to some obtained in this section have been obtained independently by Culver [8] in considering the approximation of a linear time-varying model by an autonomous model of equal order.

and algebraic manipulations yield the following expression for the minimum functional -

$$J_1^0 = \|y'\|^2_{V_y} \tag{50}$$

(b) For the third configuration it is reasonable to consider the functional

$$J_3 = \|y' - x_3(t_1)\|^2_F + \int_0^{t_1} \|u_3(t)\|^2_H dt \tag{51}$$

with the assumption: $x_3(0) = 0$. Then

$$u_3^0(t) = H^{-1}R_r^t(t_1,t)V_r y' \tag{52}$$

$$y' - x_3(t_1) = [I_{n-m} + W_r(t_1)F]^{-1}y' \tag{53}$$

and

$$J_3^0 = \|y'\|^2_{V_r} \tag{54}$$

where

$$V_r = F[I_{n-m} + W_r(t_1)F]^{-1} \tag{55}$$

(c) Now consider the second configuration, in which the plant is subject to $u_3^0(t)$. It may be shown that the target-miss for trajectory 2 is

$$y' - Cx_2(t_1) = [I_{n-m} + (W_r - W_c)F][I_{n-m} + W_r F]^{-1}y' \tag{56}$$

Algebraic manipulations yield

$$J_2 = y'^t[V_r^t W_r V_r + (I_{n-m} - W_c V_r)^t F(I_{n-m} - W_c V_r)]y' \tag{57}$$

and,

$$\int_0^{t_1} \|u_1^0(t) - u_3^0(t)\|^2_H dt = y'^t[V_y^t W_y V_y - 2V_y^t W_c V_r + V_r^t W_r V_r]y' \tag{58}$$

(d) When $W_c = W_r$, the following expression is obtained for the target-miss in trajectory 2 by substituting W_r for W_c in (56),

$$y' - y_2(t_1) = [I_{n-m} + W_r F]^{-1}y' = y' - x_3(t_1) \tag{59}$$

Hence the target-miss in trajectory 2 is equal to the target-miss in trajectory 3.

Also, from (58),

$$\int_0^{t_1} \|u_1^0(t) - u_3^0(t)\|^2_H dt = y'^t[V_y^t W_y V_y - (2V_y - V_r)^t W_r V_r]y' \tag{60}$$

Finally, from (57)

$$J_2 = y'^t[I+FW_r]^{-1}F[I+W_r F]^{-1}y' + y'^t[I+FW_r]^{-1}FW_r F][I+W_r F]^{-1}y'$$

$$= y'^t F[I+W_r F]^{-1}y' = \|y'\|^2_{V_r} \tag{61}$$

Hence for $W_c(t_1) = W_r(t_1)$ the following important results are obtained

$$J_2 = J_3^0 = \|y'\|^2_{V_r} \tag{62}$$

$$J_2 - J_1^0 = \|y'\|^2_{V_r - V_y} \tag{63}$$

Note that, as may easily be shown, when F is nonsingular

$$V_r - V_y = (F^{-1}+W_y)^{-1}(W_y - W_r)(F^{-1}+W_r)^{-1} \tag{64}$$

By assuming that the target y' is distributed over the output space with a probability density such as that given in (34), results corresponding to those given in Section 2.1, (e) may be obtained. In particular for

$$Q = (F^{-1}+W_r) \tag{65}$$

$$M[J_2 - J_1^0] = \text{Trace}(F^{-1}+W_y)^{-1}(W_y - W_r) \tag{66}$$

This expression should be compared with that given in (18) for the error functional in the reduction process with the constraint $W_c(t_1) = W_r(t_1)$.

2.3 Closed Loop Optimal Output Regulation

The optimal feedback regulator derived for the reduced model is used to regulate the plant by operating on the latter's output.

(a) For the first configuration the functional is

$$J_1 = \|y_1(t_1)\|^2_E + \int_0^{t_1} \left[\|y_1(t)\|^2_G + \|u_1(t)\|^2_H \, dt \right] \tag{67}$$

and it is given that $x(0) = x'$. H is assumed to be positive-definite and E and G are assumed to be positive semi-definite.

The solution to the optimal output regulation problem is well known [9] and it will suffice to state the relevant results.

$$u_1^0(t) = - H^{-1}B^t K_1(t,t_1)x(t) \tag{68}$$

$K_1(t,t_1)$ is the solution of the matrix Riccati equation,

$$- \frac{d}{dt} K_1 = F_1^t K_1 + K_1 F_1 + Q_1 \tag{69}$$

where

$$F_1(t_1,t) = [A - BH^{-1}B^t K_1(t,t_1)], \tag{70}$$

$$K_1(t_1,t_1) = C^t E C \tag{71}$$

and

$$Q_1(t,t_1) = K_1(t,t_1)BH^{-1}B^t K_1(t,t_1) + C^t G C \tag{72}$$

$$K_1(t,t_1) = \int_0^{t_1} \phi^t_{F_1}(\tau,t) Q_1(t,t_1) \phi_{F_1}(\tau,t) d\tau \tag{73}$$

where $\phi_{F_1}(t,t_0)$ is the transition matrix of F_1.

Finally,

$$J_1^0 = x'^t K_1(t_1)x' \tag{74}$$

where

$$K_1(t_1) = K_1(0,t_1)$$

(b) For the third configuration it is reasonable to consider the functional

$$J_3 = \|x_r(t_1)\|_E^2 + \int_0^{t_1} \left[\|x_r'(t_1)\|_G^2 + \|u_3(t)\|_H^2 \right] dt \tag{75}$$

Then, $$u_3^0(t) = - H^{-1}B_r^t K_3(t,t_1)x_r(t) \tag{76}$$

$K_3(t,t_1)$ is the solution of the Riccati equation

$$- \frac{d}{dt} K_3 = F_3^t K_3 + K_3 F_3 + Q_3 \tag{77}$$

with the boundary condition

$$K_3(t_1,t_1) = E \tag{78}$$

and $$F_3(t,t_1) = [A_r - B_r H^{-1} B_r^t K_3(t,t_1)] \tag{79}$$

$$Q_3(t,t_1) = K_3(t,t_1)B_r H^{-1} B_r^t K_3(t,t_1) + G \tag{80}$$

Also, $$J_3^0 = x_r'^t K_3(t_1)x_r' \tag{81}$$

where, $$x_r(0) = x_r' \quad \text{and} \quad K_3(t_1) = K_3(0,t_1)$$

(c) Now consider the second configuration; the central result is the derivation of J_2. In this configuration the optimal feedback controller for X_r is used to control X by operating on the output of X.

It follows from (76) that

$$u_2(t) = - H^{-1}B_r^t K_3(t,t_1)Cx_2(t) \tag{82}$$

The response of the system thus controlled is the solution of

$$\frac{d}{dt} x_2(t) = F_2(t,t_1)x_2(t) \tag{83}$$

where $$F_2(t,t_1) = [A - BH^{-1}B_r^t K_3(t,t_1)C] \tag{84}$$

and $$x_2(0) = x'$$

The transition matrix of F_2 is denoted by $\phi_{F2}(t,t_0)$. Also $\phi_{F2}(t,0)$ is denoted by $\phi_{F2}(t)$. Now it follows from straightforward substitution that

$$J_2 = \|y_2(t_1)\|_E^2 + \int_0^{t_1} \left[\|y_2(t)\|_G^2 + \|u_2(t)\|_H^2 \right] dt \tag{85}$$

$$= x'^t \left[\int_0^{t_1} \phi_{F2}^t(t) Q_2(t,t_1) \phi_{F2}(t) dt + \phi_{F2}^t(t_1) C^t E C \phi_{F2}(t_1) \right] x' \tag{86}$$

where $$Q_2(t,t_1) = C^t K_3(t,t_1) B_r H^{-1} B_r^t K_3(t,t_1) C + C^t G C \tag{87}$$

The bracketed expression in (86) is the solution of a matrix Riccati equation. In fact,

$$J_2 = x'^t K_2(t_1)x' \tag{88}$$

where $K_2(t_1) = K_2(0,t_1)$ and $K_2(t,t_1)$ is the solution of

$$- \frac{d}{dt} K_2 = F_2^t K_2 + K_2 F_2 + Q_2 \tag{89}$$

with the boundary condition $K(t_1,t_1) = C^t E C$.

Clearly $$J_2 - J_1^0 = x'^t[K_2(t_1) - K_1(t_1)]x' \qquad (90)$$

3. CONCLUSIONS

Three specific optimal control problems have been considered, and in each case explicit expressions have been obtained for the deterioration in performance when the plant is controlled as though it were the model. In the first two problems considered the equality $W_c(t_1) = W_r(t_1)$ has been shown to be important insofar as it implies the equality, performance-wise, of configuration 2 (plant controlled by model equations) and 3 (model controlled by model equations). Furthermore it has been shown that in the former problem the condition on the terminal point is satisfied only if the above-mentioned equality holds.

To a limited extent it has been possible to associate the deterioration in performance in model control of plant with the choice of the error functional for the problem of model reduction. This, of course, is highly desirable in that it allows a meaningful formulation of the reduction problem.

A major drawback of model control in plant-output regulation is the inability to guarantee stability. Further effort is required to delimit the class of reduced models which guarantee stability in control of this form.

REFERENCES

[1] (a) Mitra, D. "The Reduction of Complexity of Linear Dynamical Models", 1967, Report R-520, Atomic Energy Establishment, Winfrith, Dorchester, Dorset, England.
(b) "The Reduction of Complexity of Linear, Time Invariant Dynamical Systems", Fourth Congress, I.F.A.C., Warsaw, 1969.

[2] Mitra, D. "On the Reduction of Complexity of Linear Dynamical Models - Results of Computer Studies", 1967, Report R-535, Atomic Energy Establishment, Winfrith, Dorchester, Dorset, England.

[3] Mitra, D. "The Equivalence and Reduction of Linear Dynamical Systems", 1967, Ph.D. thesis, University of London.

[4] Mitra, D. "The W Matrix and the Geometry of Model Equivalence and Reduction", Proc. IEE, Vol. 116, No. 6, June 1969.

[5] Kantarovich, L. and Akrilov, G. "Functional Analysis in Normed Spaces" (Macmillan, 1964).

[6] Bellman, R. "Introduction to Matrix Analysis", (McGraw-Hill, 1955).

[7] Kalman, R. E., Ho, Y.C., and Narendra, K.S., "Controllability of Linear Dynamical Systems", Contributions to Differential Equations, 1963, Vol. 1, No. 2.

[8] Culver, W.J. "An Analytic Theory of Modeling for a Class of Minimal-Energy Control Systems (Disturbance-Free Case)", SIAM Control, 1964, Vol. 2, No. 3.

[9] Kalman, R.E. "Contributions to the Theory of Optimal Control", Bol. Soc. Mat. Mex., 1960, Vol. 5, pp. 102-119.

[10] Mitra, D. "Analytical Results on the Use of Reduced Models in the Control of Linear Dynamical Systems", Proc. IEE, Vol. 116, No. 8, August 1969.

CONJUGATE GRADIENT MINIMIZATION TECHNIQUE IN PROCESS OPTIMIZATION

S. ATLURI
Canadian Westinghouse Company Limited
Hamilton, Ontario, Canada

ABSTRACT

The conjugate gradient method is considered for process optimization problems for unconstrained optime. In particular we consider a function of n variables for which the gradient and function value can be calculated analytically, as is the case in some process optimization problems.

The basic algorthim of Fletcher and Reeves is extended so as to obtain a suitable computational procedure which can be easily programmed, with relatively small core requirements. It takes a small amount of time to execute the program in a real-time (on-line) process control computer system. Major details of computational procedure are described.

An application of the above method in an on-line control of blending of various raw materials in a cement making process is discussed.

INTRODUCTION

In an on-line computer system, the digital computer most often performs monitoring, data processing and optimal control functions. In this paper the optimal control phase is considered. In the optimal control phase the computer attempts to optimize the system by determining optimum values of the variables in the system satisfying the constraints specified by the operator, which are usually sets of set points. The computer optimizes the system with speed and accuracy, which the operator may accomplish with very great difficulty or in some processes may not be able to do at all. The choice of the performance index usually regulates the resulting control, in the optimum design of control systems. Here systems with performance index of the type $F = \sum_{m=1}^{n} G_m (e_m)^2$, where e_m are system error functions from set points in the time interval t_{i+1} are considered. This type of performance index is common in computer control system.

To optimize systems with the above type of performance index, Bellman's dynamic programming can be applied. However, the main drawback is that core requirements are large for even relatively simple problems. Also the time required for execution of the program is large. With the current availability of several rapidly convergent finite dimensional unconstrained minimization techniques, it is apparent that these methods can be successfully applied to problems of the above nature.

The method of conjugate gradients minimization technique is applied to the optimal control problem, because of its rapid convergence and small core requirements and small execution time. The superiority of the method has been established by several authors in recent years (1), (2), (3) and (4). Lasdon, Mitter and Warren (5) applied this method to quadratic performance criteria in optimal control systems. Recently Sinha and Iqleem (6) applied the same to systems based on ITSE (Integral of Time and Squared Error) performance index.

The basic algorthim of Fletcher and Reeves is extended so as to obtain a computational algorthim. Major computational details are given.

The application of the method in an on-line computer control system of blending several raw materials in cement making process is described.

CONJUGATE GRADIENT METHOD

Consider a system with performance index of the type,

$$F(x) = \sum_{i=1}^{n} G_i \left[e_i(x) \right]^2 \quad . \tag{1}$$

In equation (1), function F (x) of n independent variables and gradient g (x) can be calculated at any point x, analytically. e_i (x) are system error functions. G_i are the error weighing factors, determined according to the relative importance of the errors in the system. The problem now is to minimize F (x), thus minimizing the errors in the system. These errors are normally deviations from set points of constraints to be met.

Algorthim and Computational Procedure:

Let j = Counter for number iterations
k = Counter for number of steps

An arbitrary start point for $X_{i,j,k}$ is defined. Using equations (2) and (3), F (j,k) and $g_{i,j}$ are calculated.

$$F(j,k) = F(X_{i,j,k}) \tag{2}$$

$$g_{i,j} = \frac{\partial F}{\partial X_{i,j,k}} = g(X_{i,j}) \tag{3}$$

The maximum of gradients $g_{i,j}$ is calculated and stored in g_{max}, in the computer. For the first iteration j = 1, set

$$Cg_{i,j} = -g_{i,j} \quad . \tag{4}$$

The step size is now determined so as to minimize F (j,k) on the line through $X_{i,j,k}$, in the direction $Cg_{i,j}$. The step size is given as in equation (5).

$$STP_i = \frac{Cg_{i,j}}{g_{max}} \cdot \alpha \cdot ST. \qquad (5)$$

The value of Δ ST is set at some arbitrary value (0.0001). Initially value α is set at 1. Now the step counter k is incremented and new values of $X_{i,j,k}$ are determined as in equation (6).

$$X_{i,j,k} = X_{i,j,k} + STP_i \qquad (6)$$

The function value F $(X_{i,j,k})$ is determined using the new values of $X_{i,j,k}$. To find that the iteration is proceeding in the direction of minimum a check of $F_{j,k} \leqslant F_{j,k-1}$ is made. If the check is false, α is changed to -1 and the above procedure is repeated until the check is true. If the check is true α is changed 2, thus proceeding with twice step size towards the minimum on this line.

To reduce the number of steps, if the step counter is greater than 2, linear interpolation is used to determine interpolated values $X^1_{i,j,k}$. The search is continued with these interpolated values until a minimum value of $F_{j,k}$ is found.

Once the minimum point is reached for the first iteration, along $Cg_{i,j}$ the step counter k is reset. Gradients $g_{i,j}$ are calculated using equation (3), using values $X_{i,j,k}$ during the last iteration at minimum F $(X_{i,j,k})$. New directions of search are found as in equations (7) and (8).

$$\beta_{i,j} = g^2_{i,j} \; / \; g^2_{i,j-1} \qquad (7)$$

$$Cg_{i,j} = g_{i,j} + \beta_{i,j} \cdot Cg_{i,j-1} \qquad (8)$$

The step counter k is incremented and the correct path towards minimum along the new direction of search $Cg_{i,j}$ is found. Search is continued along this path using the above procedure until minimum point is found. Each time $X_{i,j,k}$ is incremented by the step size given in equation (5).

The method does not attempt to locate a precise minimum of F $(X_{i,j,k})$ in any given direction. The method recalculates the gradients only when a local minimum is found along the given direction, thus reducing the number of gradient and function evaluations, thereby reducing the program execution time.

True convergence is reached only when all $g_{i,j}$ vanish. However, this is impossible to realize in practice because of the round of errors. Iteration is continued each time updating the gradients and directions of search using equations (7) and (8) till no further reduction in $F_{j,k}$ is found. Arbitrarly the reduction δ is set at 1×10^{-5}.

The choice of start point is in principle arbitrary. In a practical system, the start point is taken as the value of $X_{i,j,k}$ determined during the previous time interval. This ensures the iteration in reaching the minimum point in as few iterations as possible. Thus when applied in an on-line control system, the method optimizes the system in a fewer number of iterations with equally less number of steps.

The flow chart in Fig. 1 describes the above computational procedure in detail.

The program takes about 350 core locations in computer. This usually varies depending on the number of variables controlled. The execution time normally depends mainly upon the number of steps and iterations taken in achieving the minimum value of the performance index, and also on the addition and multiply times of the computer used. The program uses floating point arithmetic.

APPLICATIONS

The method can be applied successfully for many process optimization problems. Examples of such problems include 1) optimization of blending of several raw materials in pulp and paper, mining and cement industries and 2) optimization of distillation column - condenser system and determination of optimum kinetic and heat transfer constants for a catalytic polymerization reaction in a polyethylene plant, (3) In computer applications to power industry, particularly in economic dispatch problems, in computing optimum assignments of received load to various generating units.

Applications of the Method in Blending Process:

The following application concerns in using the above method in optimal control phase of a real-time computer system used to control blending process in cement industry.

In any cement manufacturing process blending of several raw-materials involves, determining optimum proportions of raw materials to achieve the desired chemical composition of clinker.

The chemical composition of clinker is specified in terms of percentages of compounds. Raw material proportions are determined such that the deviations from set points of the clinker compounds are minimized during the next time interval. For a system of n + 1 raw materials there will be n independent variables, because of the fact that the sum of raw material proportions is equal to 1. Also to have one and only one minima of index, the clinker compounds should also be equal to n. Thus for a n-independent variable system the blending process error function E_m is given as in equation (9).

$$E_m = (M_m)_o - (M_m)_{t_i} - (M_m)_{t_{i+1}}, \quad m=1,2,---n \tag{9}$$

where $(M_m)_o$ = clinker compounds set points

$(M_m)_{t_i}$ = clinker compounds computed by an on-line computer, with inputs from a x-ray analyzer at time t_i using Bogue (7) equations.

$(\Delta M_m)_{t_{i+1}}$ represents the mathematical model of the blending system given in Table 1. These are difference equations with changes in percent clinker compounds during the time interval t_i to t_{i+1}.

TABLE 1

Mathematical Model of Blending Process

ΔM_m	
ΔM_1	$= \sum_{m=1}^{n} \alpha_m \Delta X_m$
ΔM_2	$= \sum_{m=1}^{n} \beta_m \Delta X_m$
ΔM_3	$= \sum_{m=1}^{n} \gamma_m \Delta X_m$
ΔM_4	$= \sum_{m=1}^{n} \delta_m \Delta X_m \Big/ \sum_{m=1}^{n} S_m \Delta X_m$
---	---
ΔM_n	-----

In Table 1, α_m, β_m, γ_m, δ_m, S_m are different functions of individual raw material oxide compositions. ΔX_m are the changes in raw material proportions which are to be found. To achieve this in a optimum way, the performance index is defined as in equation (10).

$$F = \sum_{m=1}^{n} G_m E_m^2 \tag{10}$$

where G_m are deviation weighing factors.

The conugate gradient method is used to minimize the function F and in doing so the optimum values of X_m are found, thereby optimizing the system.

The start points for ΔX_m are taken as the values calculated during the previous time interval, which are stored in the computer. Typical results for a system of 4 and 5 raw materials are shown in Table 2 and Table 3, noting that for n+1 raw material system there will be n independent variables (raw material proportions). ΔST is taken as 0.00125. The program is terminated when no further reduction of $\delta = 1 \times 10^{-5}$ is achieved in the function value. The convergence is reached in 28 iterations in 4 variable system and in 22 iterations in 3 variable system.

In a similar way the method can be applied to blending processes in mining and pulp and paper industries.

CONCLUSIONS

The method of conjugate gradients seems very effective in applications to process optimization problems. The algorthim can be easily programmed. It takes a small amount of core in computer. Execution time is relatively short. It requires a small number of iterations, with an equally small number of steps to converge to a minimum point.

The method can be extended to problems with inequality constraints. It is also apparent that other existing finite dimensional minimization techniques can be similarly extended in application to practical problems, particularly the method of parallel tangents seems promising.

REFERENCES

(1.) R. Fletcher and C.M. Reeves: "Function Minimization by Conjugate Gradients", British Computer Journal, July 1964.

(2.) M.R. Hestenes and E. Stiefel: "Method of Conjugate Gradients for Solving Linear Systems", J. Research NBS, Vol. 49, 1952.

(3.) F.S. Beckman: "The Solution of Linear Equations by the Conjugate Gradient Method", Mathematical Methods for Digital Computers, A. Ralston, H.S. Wilf (Eds.), Wiley.

(4.) M.J. Box: "A Comparison of Several Current Optimization Methods and the Use of Transformations in Constrained Problems", The Computer Journal, Vol. 9, 1966.

(5.) L.S. Lasdon, S.K. Mitter and A.D. Warren: "The Conjugate Gradient Method for Optimal Control Problems", IEEE Transc. Ac. April, 1967.

(6.) N.K. Sinha, M.J. Iqleem: "Adaptive Control without Complete Plant Identification", proceedings 6th Allerton Conference, Urbana, Illinois, PP. 201, Oct. 1968.

(7.) Bogue: "The Chemistry of Portland Cement", Reinhold, 1947.

(8.) S.R. Atluri and W.B. Duffey: "Digital Control of Dry Material Blending", Instruments and Control Systems, Vol. 41, Nov. 1968.

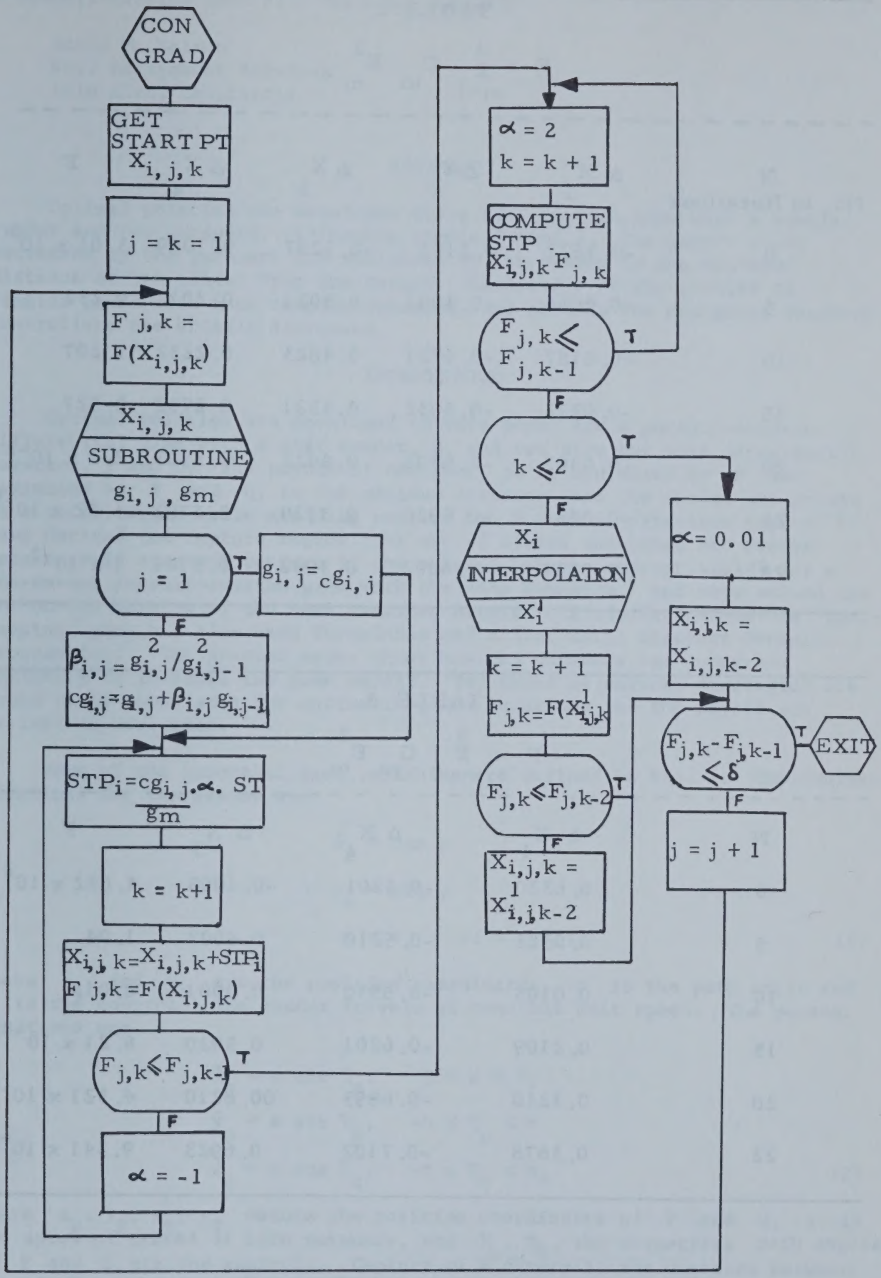


FIG 1. CONGUGATE GRADIENT OPTIMIZATION PROGRAM

TABLE 2

$$F = \sum_{m=1}^{4} G_m E_m^2$$

N No. of Iterations	ΔX_1	ΔX_2	ΔX_3	ΔX_4	F
0	-0.3621	0.2101	-0.1207	0.1009	3.01×10^{3}
5	-0.0094	-0.4001	0.5021	0.2012	9.232
10	-0.0187	-0.4721	0.4823	0.2232	2.207
15	-0.0272	-0.5032	0.4221	0.2732	0.527
20	-0.0302	-0.5103	0.4022	0.3103	2.2×10^{-2}
25	-0.0456	-0.5920	0.3320	0.4702	4.32×10^{-5}
28	-0.0531	-0.6002	0.3002	0.5104	1×10^{-7}

TABLE 3

$$F = \sum_{m=1}^{3} G_m E_m^2$$

N	ΔX_1	ΔX_2	ΔX_3	F
0	0.6320	-0.3201	-0.1003	4.032×10^{2}
5	0.0321	-0.5210	0.4002	1.04
10	0.0103	-0.5530	0.5001	0.013
15	0.2109	-0.6201	0.5920	8.21×10^{-3}
20	0.3210	-0.6893	00.6210	4.321×10^{-5}
22	0.3678	-0.7102	0.6923	9.341×10^{-6}

A PURSUIT-EVASION GAME WITH MULTIPLE PURSUERS

David M. Salmon
Wolf Management Services
Palo Alto, California

ABSTRACT

Optimal policies are developed for a differential game with a single evader and two pursuers, all having simple dynamics. The payoff to be maximized by the pursuers and minimized by the evader is the minimum distance of the evader from the target. Extensions of the results to similar games with more complex dynamics, and methods for designing feedback controllers are briefly discussed.

INTRODUCTION

Optimal policies are developed in this paper for a pursuit-evasion differential game with a fast evader, E, and two slow but more maneuverable pursuers, P and Q. The payoff of the game, to be minimized by E and maximized by P and Q, is the minimum distance that the evader can attain to a point target while avoiding capture by P or Q. Previous papers[1-3*] have derived the capture region (the set of states such that the evader cannot avoid eventual capture if the pursuer behaves appropriately) for a one-on-one pursuit-evasion game with the same dynamics,[1] and have solved the one-on-one version of the game examined herein.[2] A similar one-on-one "goal-keeping" game has also been formulated and solved using discrete dynamic programming.[3] The present paper shows how two pursuers can cooperate optimally to increase the game payoff. Two types of pursuer strategies are found to be candidates for optimality, and equations for the payoff are derived in each case.

Some of the important game variables are defined in Fig. 1. The dynamic equations for the evader are

$$\dot{x}_e = \cos\varphi$$

$$\dot{y}_e = \sin\varphi$$

$$\dot{\varphi} = v, \qquad -1 \le v \le 1 \tag{1}$$

where x_e and y_e are the position coordinates, φ is the path angle and v is the control. The evader travels at constant unit speed. The pursuer equations are

$$\dot{x}_p = a\cos\eta_p, \qquad 0 < a < 1,$$

$$\dot{y}_p = a\sin\eta_p, \qquad -\pi \le \eta_p \le \pi$$

$$\dot{x}_q = a\cos\eta_q, \qquad -\pi \le \eta_q \le \pi, \tag{2}$$

where x_p, y_p, x_q, y_q denote the position coordinates of P and Q, a is the speed of travel of both pursuers, and η_p, η_q, the respective path angles of P and Q, are the controls. Capture of E occurs if the distance between P and E or between Q and E becomes equal to c, where $c < 1$. Initially the evader is a long way from the target which is at the origin ($x = 0$, $y = 0$). The pursuers are permitted to choose initial conditions in

the vicinity of the target. The payoff to be minimized by E and maximized by P and Q is the minimum distance to the origin attained by E.

Critical Terminal Conditions and Optimal Trajectories

The key to solution of this differential game is the recognition that, at some instant of time in an optimal play, E must be on the verge of capture by P, or by Q, or by P and Q. If there is no such instant of time it follows that either the minimum distance from the evader to the target is zero (shown to be false in the following) or the evader could have reduced the minimum distance to the target without capture. The conditions corresponding to E being on the verge of capture by one pursuer, and by two pursuers, are derived below. The necessary conditions for optimality are then employed to generate optimal trajectories for the two cases.

Consider first the one-on-one situation with a pursuer, say P, very close to the boundary of the capture circle about E, i.e.

$$\begin{aligned} x_p - x_e &= c \cos s_1 \\ y_p - y_e &= c \sin s_1 \end{aligned} \quad , \tag{3}$$

where s_1 is the angle of the line joining E to P. The velocity V_R of E relative to P along the inward directed normal to the capture circle,

$$(\nu_1, \nu_2) = (-\cos s_1, -\sin s_1) \quad , \tag{4}$$

is given by

$$V_R = [\nu_1 \dot{x}_e + \nu_2 \dot{y}_e - \nu_1 \dot{x}_p - \nu_1 \dot{y}_p] \quad .$$

To resist capture the evader should maximize V_R, and to achieve capture of the evader the pursuer should minimize V_R. Thus the evader is able to resist capture only if

$$\max_{\nu} \min_{\eta_p} [\nu_1 \dot{x}_e + \nu_2 \dot{y}_e - \nu_1 \dot{x}_p - \nu_2 \dot{x}_p] > 0 \quad . \tag{5}$$

Substituting Eqs. (1), (2), and (4) into Eq. (5), and solving,

$$\cos (s_1 - \varphi) < -a$$

i.e.,

$$\varphi - \frac{\pi}{2} - \sin^{-1} a < s_1 < \varphi + \frac{\pi}{2} + \sin^{-1} a \quad . \tag{6}$$

Similarly, P can guarantee capture if the inequalities in (6) are reversed. Therefore, the critical conditions where E is on the verge of capture correspond to

$$s_1 = \varphi + \frac{\pi}{2} + \sin^{-1} a$$

and

$$s_1 = \varphi - \frac{\pi}{2} - \sin^{-1} a \quad . \tag{7}$$

With pursuers P and Q at positions such that the states of P satisfy Eqs. (5) and the states of Q satisfy

$$x_q - x_e = c \cos s_2$$
$$y_q - y_e = c \sin s_2 \quad , \tag{8}$$

the conditions for E to resist capture are

$$\max_{v} \min \left\{ \min_{\eta_p} [\nu_1 \dot{x}_e + \nu_2 \dot{y}_e - \nu_1 \dot{x}_p - \nu_2 \dot{y}_p], \min_{\eta_p} [\nu_3 \dot{x}_e + \nu_4 \dot{y}_e - \nu_3 \dot{x}_p - \nu_4 \dot{y}_p] \right\} > 0 \tag{9}$$

which reduces to

$$\cos (s_1 - \varphi) < - a$$

and

$$\cos (s_2 - \varphi) < - a \quad . \tag{10}$$

The critical conditions, where E is on the verge of capture by P, by Q, or by P and Q, correspond respectively to

$$s_1 = \begin{cases} \varphi + \frac{\pi}{2} + \sin^{-1} a \\ \varphi - \frac{\pi}{2} - \sin^{-1} a \end{cases} ,$$

or

$$s_2 = \begin{cases} \varphi + \frac{\pi}{2} + \sin^{-1} a \\ \varphi - \frac{\pi}{2} - \sin^{-1} a \end{cases} ,$$

or

$$\begin{cases} s_1 = \varphi + \frac{\pi}{2} + \sin^{-1} a \text{ and } s_2 = \varphi - \frac{\pi}{2} - \sin^{-1} a \\ \text{or } s_1 = \varphi - \frac{\pi}{2} - \sin^{-1} a \text{ and } s_2 = \varphi + \frac{\pi}{2} + \sin^{-1} a \end{cases} . \tag{11}$$

Let $\bar{t}$ be the latest instant of time at which E is on the verge of capture. After $\bar{t}$, the optimal policy for E is to make a maximum turn towards the origin. Accordingly, as in Ref. 2, the payoff, D, is simply a function of the state of E at time $\bar{t}$,

$$D = D[x_e(\bar{t}), y_e(\bar{t}), \varphi(\bar{t})] \quad . \tag{12}$$

(The detailed form of D is not required for the analysis to follow.) Thus the game is equivalent to a game ending at $\bar{t}$ with payoff D.

The necessary conditions[4] for the optimal trajectories before time $\bar{t}$, for the case where E is on the verge of capture by both P and Q at $\bar{t}$, follow. The Hamiltonian is given by

$$H = \lambda_1 \cos \varphi + \lambda_2 \sin \varphi + \lambda_3 v + \lambda_4 a \cos \eta_p + \lambda_5 a \sin \eta_p + \lambda_6 a \cos \eta_q + \lambda_7 a \sin \eta_q \quad . \tag{13}$$

The adjoint variables are defined by

$$
\begin{aligned}
\dot{\lambda}_1 &= -H_{x_e} = 0 && , \quad \lambda_1(\bar{t}) = \Phi_{x_e}(\bar{t}) \\
\dot{\lambda}_2 &= -H_{y_e} = 0 && , \quad \lambda_2(\bar{t}) = \Phi_{y_e}(\bar{t}) \\
\dot{\lambda}_3 &= -H_{\varphi} = \lambda_1 \sin\varphi - \lambda_2 \cos\varphi && , \quad \lambda_3(\bar{t}) = \Phi_{\varphi}(\bar{t}) \\
\dot{\lambda}_4 &= -H_{x_p} = 0 && , \quad \lambda_4(\bar{t}) = \Phi_{x_p}(\bar{t}) \\
\dot{\lambda}_5 &= -H_{y_p} = 0 && , \quad \lambda_5(\bar{t}) = \Phi_{y_p}(\bar{t}) \\
\dot{\lambda}_6 &= -H_{x_q} = 0 && , \quad \lambda_6(\bar{t}) = \Phi_{x_q}(\bar{t}) \\
\dot{\lambda}_7 &= -H_{y_q} = 0 && , \quad \lambda_7(\bar{t}) = \Phi_{y_q}(\bar{t}) \quad , \qquad (14)
\end{aligned}
$$

where

$$
\Phi = D + \mu_1[(x_e - x_p)^2 + (y_e - y_p)^2 - c^2] + \mu_2[(x_e - x_q)^2 + (y_e - y_q)^2 - c^2] \quad .
$$

The optimal controls $\bar{v}$, $\bar{\eta}_p$, $\bar{\eta}_q$, must satisfy

$$
\min_{\eta_p, \eta_q} \max_{v} H(\eta_p, \eta_q, v) = H(\bar{\eta}_p, \bar{\eta}_q, \bar{v}) \quad . \qquad (15)
$$

From Eqs. (13), (14), and (15), and using the terminal conditions of Eqs. (5) and (8),

$$
\begin{aligned}
\bar{\eta}_p(t) &= \pi + s_1 \\
\bar{\eta}_q(t) &= \pi + s_2 \\
\bar{v} &= \operatorname{sgn} \lambda_3 \quad \text{if} \quad \lambda_3 \neq 0 \\
\lambda_3(\bar{t}) &= \operatorname{sgn} D_{\varphi}(\bar{t}) \quad . \qquad (16)
\end{aligned}
$$

Using, from Eq. (11),

$$
s_1 = \varphi + \frac{\pi}{2} + \sin^{-1} a
$$

and

$$
s_2 = \varphi - \frac{\pi}{2} - \sin^{-1} a \quad ,
$$

the optimal controls for P and Q are

$$
\begin{aligned}
\bar{\eta}_p(t) &= \varphi + \frac{3\pi}{2} + \sin^{-1} a \quad , \\
\bar{\eta}_q(t) &= \varphi + \frac{\pi}{2} - \sin^{-1} a \quad . \qquad (17)
\end{aligned}
$$

Note that there are only three candidates for an optimal control for v:

$$
\begin{aligned}
\bar{v} &= +1 \quad \text{for} \quad \lambda_3 < 0 \\
\bar{v} &= -1 \quad \text{for} \quad \lambda_3 > 0 \\
\bar{v} &= 0 \quad \text{for} \quad \lambda_3 \equiv 0 \quad \text{[i.e., if } \lambda_3(t) = 0 \text{ on an interval of time.]} \qquad (18)
\end{aligned}
$$

The last result follows from

$$\lambda_3 \equiv 0 \Rightarrow \dot{\lambda} = 0 \Rightarrow \varphi = \text{const} \Rightarrow \bar{v} = 0 \quad ,$$

using Eqs. (14) and (1).

A similar analysis for one pursuer, say P, on the verge of capture at $\bar{t}$ leads to the results

$$\bar{\eta}_p(t) = \pi + s_1$$

$$\bar{v} = \text{sgn}\, \lambda_3$$

$$\bar{v}(\bar{t}) = \text{sgn}\, D_\varphi(\bar{t}) \quad . \tag{19}$$

The implications of Eqs. (17), (18), and (19), are summarized in Fig. 2. If, at time $\bar{t}$, the evader is at $E_{\bar{t}}$, either or both of the pursuers must be at either or both of $P_{\bar{t}}$ and $Q_{\bar{t}}$. The evader must follow one of the three trajectories $E^1 E_{\bar{t}}$, $E^2 E_{\bar{t}}$, or $E^3 E_{\bar{t}}$. No other trajectories can be optimal.

Simultaneous Pursuer Policy

It is now necessary to combine the segments of optimal trajectories shown in Fig. 2 and determine their positioning on the x - y plane. Figure 3 shows one combination, called the simultaneous pursuer policy. At time t_o, the evader is traveling directly towards the origin (the evader is shown at E_{t_o} on the y-axis in Fig. 3) while the two pursuers are at the corresponding positions P_{t_o} and Q_{t_o}. To avoid capture the evader must turn either to the left or the right and narrowly evade P at $P_{\bar{t}}$, or Q at $Q_{\bar{t}}$. Finally, E turns towards the origin, achieving the payoff $\bar{t} D^*$ shown. The alternative trajectory for E to travel directly to $E^3_{\bar{t}}$ will lead to capture of E providing P and Q are at the correct initial positions P_{t_o}, Q_{t_o}, and follow the trajectories $P_{t_o} P^2_{\bar{t}}$, $Q_{t_o} Q^2_{\bar{t}}$. Note that the conditions at time $\bar{t}$ in Fig. 3 agree with the necessary conditions; i.e., $D_\varphi(\bar{t})$ is positive at $E^1_{\bar{t}}$, negative at $E^2_{\bar{t}}$, and zero at $E^3_{\bar{t}}$, thus implying via Eqs. (19) and (1) that E is turning hard to the left before reaching $E^1_{\bar{t}}$, hard to the right before $E^2_{\bar{t}}$, and allowing for a singular arc before reaching $E^3_{\bar{t}}$. Also, it is evident from the geometry of Fig. 3 that E can reduce the minimum distance to the target if the pursuers deviate from the above policies and that the pursuers can either capture E or increase the minimum distance of E from the target if E deviates from the above policy. Thus the strategies described constitute a solution to the game.

However, the described trajectories are not optimal if the optimal payoff, D^*, becomes greater than 1.0. In this case, the same relative trajectories of E, P and Q, are candidates for optimality but it is now to the advantage of E to shift the trajectories on the x - y plane by not traveling directly towards the origin at time t_o. The game may be solved for given values of a and c simply by determining graphically how E, and P and Q, should position the relative trajectories shown in Fig. 3. The analytical expression for the optimal payoff becomes complex in this case.

Equations for the optimal payoff D^* for $D^* < 1.0$ can be derived from the geometry of Fig. 3. Let t_1 be the time for E to travel from

E_{t_o} to $E_{\bar{t}}^{3}$, and t_2 be the time for E to travel from E_{t_o} to $E_{\bar{t}}^{2}$. In the same time P and Q will travel distances of at_1 and at_2, respectively.

$$x_1 = (c + at_1)\cos(\sin^{-1} a)$$

$$= (1 - \cos t_2) - (c + at_2)\cos(\cos^{-1} a - \tfrac{\pi}{2} + t_2) \qquad (20)$$

$$y_1 = t_1 - (c + at_1)\sin(\sin^{-1} a)$$

$$= \sin t_2 + (c + at_2)\sin(\cos^{-1} a - \tfrac{\pi}{2} + t_2) \qquad (21)$$

Solving Eqs. (20) and (21) for t_1,

$$t_1 = \frac{1}{a\sqrt{1-a^2}}\left[(1 - \cos t_2) - (c + at_2)\cos(t_2 - \sin^{-1} a)\right] - \frac{c}{a} \qquad (22)$$

and

$$\frac{a}{\sqrt{1-a^2}}\left[\sin t_2 + (c + at_2)\sin(t_2 - \sin^{-1} a) + \frac{c}{a}\right] = 1 - \cos t_2$$

$$- (c + at_2)\cos(t_2 - \sin^{-1} a) \quad . \qquad (23)$$

Let t_2^* be the smallest positive solution of Eq. (23) and t_1^* the corresponding value of t_1. Then the optimal payoff is

$$D^* = 2 - 2\cos t_2^* \quad . \qquad (24)$$

The position of E at time t_o is given by

$$x_e(t_o) = 0$$

$$y_e(t_o) = 2\sin t_2^* \quad , \qquad (25)$$

and the positions of P and Q at time t_o are given by

$$x_q(t_o) = -x_1(t_1^*) \quad , \quad y_q(t_o) = 2\sin t_2^* - y_1(t_1^*)$$

$$x_p(t_o) = +x_1(t_1^*) \quad , \quad y_p(t_o) = 2\sin t_2^* - y_1(t_1^*) \quad . \qquad (26)$$

Sequential Pursuer Policy

An alternative solution, called the sequential pursuer policy is shown in Fig. 4. At time t_o the evader at E_{t_o} is forced by the pursuer at P_{t_o} to turn either to the left or the right. Once E has chosen a direction, both pursuers travel in that direction; P travels to $P_{\hat{t}}$ and Q to $Q_{\hat{t}+t'}$. At time $\hat{t}$, the evader narrowly evades P and turns back. At time $\hat{t}+t'$, the evader at $E_{\hat{t}+t'}$ is forced by the pursuer at $Q_{\hat{t}+t'}$ to turn outward again if capture is to be avoided. The evader narrowly evades Q at time $\bar{t}$ and finally turns towards the origin.

It is clear from the geometry of Fig. 4 that these trajectories, which satisfy the necessary conditions for optimality, also form a candidate solution for the game; i.e., if the pursuers deviate from the given policy the evader can reduce the minimum distance to the target, and if the evader deviates from the given policy the pursuers can either capture the evader

or increase the minimum distance. However, the policy of Fig. 4 will only be optimal if the payoff is less than 1.0 for the same reasons as with the simultaneous pursuer policy.

Equations for the optimal payoff D^* for $D^* < 1.0$ can be derived from the geometry of Fig. 4.

The values of $\hat{t}$ and ℓ_1 are given by[2]

$$\ell_1 = \sin \hat{t} + (c + a\hat{t}) \sin (\hat{t} - \sin^{-1} a)$$

where $\hat{t}$ is the smallest positive solution of

$$(c + a\hat{t}) \cos (\hat{t} - \sin^{-1} a) + \cos \hat{t} - 1 = 0 \quad .$$

The value of t' is given by the smallest positive solution of

$$\begin{aligned} x_1 &= 1 - 2 \cos \hat{t} + \cos (t' - \hat{t}) - \ell_1 \sin (t' - \hat{t}) \\ &= a(\hat{t} + t') \quad . \end{aligned}$$

Finally the positions of the evader and pursuers at time t_o are,

$$y_e(t_o) = 2 \sin \hat{t} + 2 \sin (t' - \hat{t}) + 2 \sin (2\hat{t} - t')$$

$$y_q(t_o) = y_e(t_o) - 2 \sin \hat{t} - \sin (t' - \hat{t}) - \ell_1 \cos (t' - \hat{t})$$

$$y_p(t_o) = y_e(t_o) - \ell_1 \quad ,$$

and the optimal payoff is (for $D^* < 1.0$)

$$D^* = 2 - 2 \cos \hat{t} + 2 \cos (t' - \hat{t}) - 2 \cos (2\hat{t} - t') \quad .$$

Choice of the final policy, sequential or simultaneous, rests with the pursuers who can optimize their performance by computing D^* for the two cases and choosing the policy that yields the larger value.

CONCLUSIONS

A final step in a complete solution of the game solved in this paper is to derive a feedback controller for the pursuers and the evader. The controller should cause the participants to follow the optimal trajectories as long as the opponent does, and to take advantage of errors (when the opponent does not follow an optimal trajectory). The controller design presented in Ref. 2 for pursuer control on the boundary of and inside the capture region can be employed for both the pursuers whenever the appropriate states occur. A simple extension of the design method can also be employed to generate a feedback controller for the remaining states of interest.

The necessary conditions for a solution to the game have been employed to determine candidate trajectories for optimality, and these trajectories have been combined in two ways to determine optimal pursuer and evader strategies. It is evident that the results can be extended in a number of ways: to similar games with more complex dynamics; to games played in three-dimensional space rather than in a plane; and to the optimal control of more than two pursuers.

REFERENCES

1. Lewis B. Meier, Jr., "A New Technique for Solving Pursuit-Evasion Differential Games," IEEE Transactions on Automatic Control, Vol. AC-14 (August 1969).

2. David M. Salmon, "Policies and Controller Design for a Pursuing Vehicle," IEEE Transactions on Automatic Control, Vol. AC-14 (August 1969).

3. Paul A. Meschler, "On a Goal-Keeping Differential Game," IEEE Transactions on Automatic Control, Vol. AC-12, No. 1 (February 1967).

4. Arthur E. Bryson, Jr., and Yu-Chi Ho, Applied Optimal Control, Blaisdell Publishing Co. (1969).

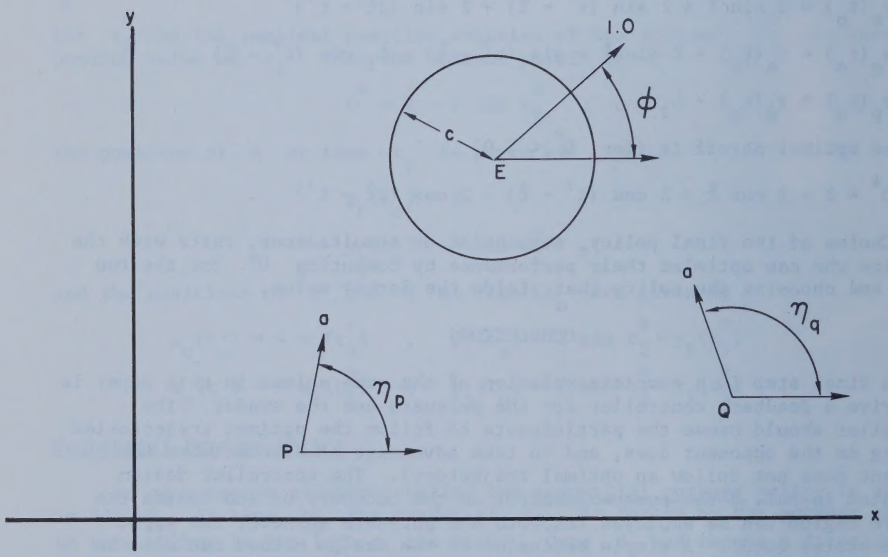


Fig. 1 Variables of the Evader, E, and the Pursuers, P and Q

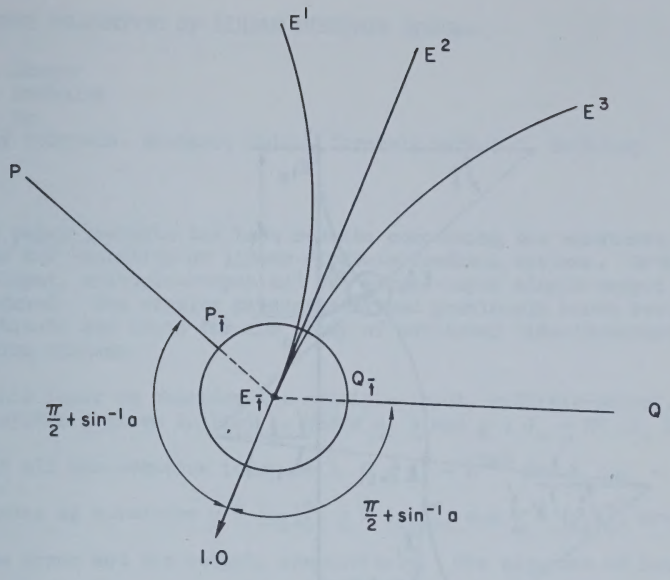


Fig. 2 Optimal Relative Trajectories Before Time $\bar{t}$

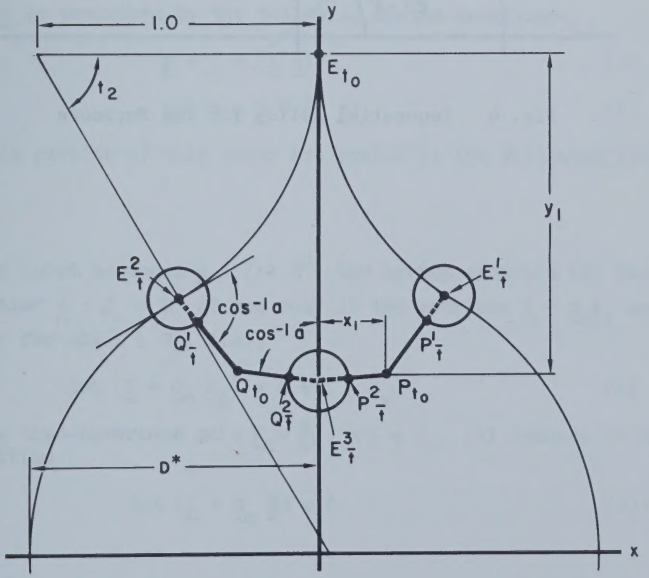


Fig. 3 Simultaneous Policy for the Pursuers

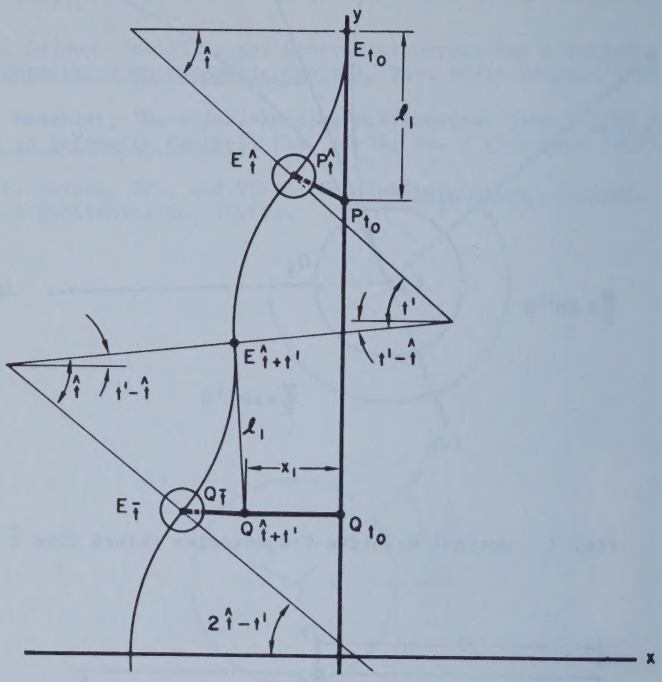


Fig. 4 Sequential Policy for the Pursuers

INPUT-OUTPUT PROPERTIES OF LINEAR DISCRETE SYSTEMS

C.A. Desoer
U.C. Berkeley
M.Y. Wu
U. of Colorado, Boulder, Colo., formerly with U.C. Berkeley

Abstract

This paper presents the best results concerning the existence, uniqueness and stability of linear discrete feedback systems. Both the multiple-input, multiple-output and the single-input single-output cases are considered. The results presented extend previously known results and they constitute key tools for the study of nonlinear time-invariant and/or time-varying systems.

In this paper we consider the multiple-input, multiple-output, linear, discrete system $\underset{\sim}{S}$ shown in Fig. 1, where $\underset{\sim}{u}$, $\underset{\sim}{e}$ and $\underset{\sim}{y} : J_+ \to R^n$ (J_+ denotes the set of all non-negative integers.), $\underset{\sim}{G} : J_+ \to R^{n\times n}$ and $\underset{\sim}{K} : J_+ \to R^{n\times n}$. The sequences of n-vectors $\underset{\sim}{u} \triangleq \{u_i\}_0^\infty$, $\underset{\sim}{e} \triangleq \{e_i\}_0^\infty$, and $\underset{\sim}{y} \triangleq \{y_i\}_0^\infty$, are the input, the error and the output, respectively. The sequence of n×n matrices $G \triangleq \{G_i\}_0^\infty$ specified the impulse response matrix which characterizes the linear dynamical subsystem. The sequence of n×n matrices $\underset{\sim}{K} \triangleq \{K_i\}_0^\infty$ specifies a memoryless time-varying gain. In contrast to a number of previous papers, we do not assume that K_i's are diagonal, or symmetric. The system $\underset{\sim}{S}$ is described by the following system equations:

$$\underset{\sim}{y} = \underset{\sim}{G} * (\underset{\sim}{K}\,\underset{\sim}{e}) \tag{1}$$

$$\underset{\sim}{e} = \underset{\sim}{u} - \underset{\sim}{G} * (\underset{\sim}{K}\,\underset{\sim}{e}) \tag{2}$$

The main results of this paper are stated in the following four theorems:

Theorem 1

For any input sequence $\underset{\sim}{u} : J_+ \to R^n$, the system equation (2) has a unique solution $\underset{\sim}{e} : J_+ \to R^n$ if and only if the matrices $\underset{\sim}{I} + \underset{\sim}{G}_0\underset{\sim}{K}_i$ are non-singular for all $i \in J_+$, i.e.,

$$\det (\underset{\sim}{I} + \underset{\sim}{G}_0\,\underset{\sim}{K}_i) \neq 0 \quad \forall\, i \in J_+. \tag{4}$$

For the time-invariant gain $\underset{\sim}{K}_i = \bar{\underset{\sim}{K}}$, $\forall i \in J_+$, (4) reduces to the single condition

$$\det (\underset{\sim}{I} + \underset{\sim}{G}_0\,\bar{\underset{\sim}{K}}) \neq 0 \tag{4}'$$

Theorem 2

Consider the multiple-input, multiple-output, linear, time-invariant discrete system $\underset{\sim}{S}$ shown in Fig. 1 which is characterized by the system

equations (1) and (2) with $\underset{\sim}{K} = \bar{\underset{\sim}{K}} \in R^{n\times n}$ being a gain matrix. Let the open-loop z-transfer function of the linear subsystem $\underset{\sim}{G}$ be of the form

$$\tilde{\underset{\sim}{G}}(z) = \underset{\sim}{R}\left(1 - z^{-1}\right)^{-1} + \sum_{i=0}^{\infty} \underset{\sim}{G}_i z^{-i} \triangleq \underset{\sim}{R}\left(1 - z^{-1}\right)^{-1} + \tilde{\underset{\sim}{G}}_{\ell}(z) \qquad (5)$$

where $\underset{\sim}{R}$ is an nxn constant matrix and $\underset{\sim}{G}_{\ell} \triangleq \{\underset{\sim}{G}_i\}_0^{\infty} = \mathcal{Z}^{-1}\{\tilde{\underset{\sim}{G}}_{\ell}(z)\} \in \ell^1_{n\times n}$.

Under these conditions, if

$$\inf_{|z| \geq 1} |\det [\underset{\sim}{I} + \tilde{\underset{\sim}{G}}(z)\, \bar{\underset{\sim}{K}}]| > 0 \qquad (6)$$

and if either $\underset{\sim}{R} = 0$ or $\underset{\sim}{R}\bar{\underset{\sim}{K}}$ is nonsingular, then

(a) the closed-loop z-transfer function (relating $\underset{\sim}{y}$ to $\underset{\sim}{u}$)

$$\tilde{\underset{\sim}{H}}(z) = [\underset{\sim}{I} + \tilde{\underset{\sim}{G}}(z)\, \bar{\underset{\sim}{K}}]^{-1}\, \tilde{\underset{\sim}{G}}(z)\, \bar{\underset{\sim}{K}} \qquad (7)$$

is analytic outside the unit disk $|z| \leq 1$, and the closed loop impulse response

$$\underset{\sim}{H} \triangleq \{\underset{\sim}{H}_i\}_0^{\infty} = \mathcal{Z}^{-1}\{\tilde{\underset{\sim}{H}}(z)\} \in \ell^1_{n\times n} \qquad (8)$$

(b) For any fixed $p \in [1,\infty]$, $\underset{\sim}{u} \in \ell^p_n$ implies that $\underset{\sim}{y} \in \ell^p_n$.

(c) $\underset{\sim}{u} \in \ell^{\infty}_n$ and $\lim_{m\to\infty} \underset{\sim}{u}_m = \underset{\sim}{0}$ implies that $\underset{\sim}{y} \in \ell^{\infty}_n$ and $\lim_{m\to\infty} \underset{\sim}{y}_m = \underset{\sim}{0}$

(d) If, in particular $\underset{\sim}{R}\bar{\underset{\sim}{K}}$ is nonsingular, then $\underset{\sim}{u} \in \ell^{\infty}_n$ and $\lim_{m\to\infty} \underset{\sim}{u}_m = \underset{\sim}{u}_{\infty}$ (a constant vector in R^n) implies that $\lim_{m\to\infty} \underset{\sim}{e}_m = \underset{\sim}{0}$ and $\lim_{m\to\infty} \underset{\sim}{y}_m = \underset{\sim}{u}_{\infty}$.

Theorem 3

Consider a multiple-input, multiple-output, linear, time-invariant discrete system such as shown in Fig. 1 and described by equations (1) and (2) with $\underset{\sim}{K} = \bar{\underset{\sim}{K}} \in R^{n\times n}$ being a time-invariant gain matrix. If

(i) $\det [\underset{\sim}{I} + \underset{\sim}{G}_0\, \bar{\underset{\sim}{K}}] \neq 0$

(ii) the open-loop impulse response sequence $\underset{\sim}{G} \triangleq \{\underset{\sim}{G}_i\}_0^{\infty}$ is exponentially bounded,

(iii) the closed-loop z-transfer function between $\underset{\sim}{u}$ and $\underset{\sim}{y}$

$$\tilde{\underset{\sim}{H}}(z) = [\underset{\sim}{I} + \tilde{\underset{\sim}{G}}(z)\, \bar{\underset{\sim}{K}}]^{-1}\, \tilde{\underset{\sim}{G}}(z)\, \bar{\underset{\sim}{K}} \qquad (9)$$

is analytic for $|z| > \rho$ where $\rho < 1$,

then the closed-loop impulse response sequence $\underset{\sim}{H} \triangleq \{\underset{\sim}{H}_i\}_0^{\infty} = \mathcal{Z}^{-1}\{\tilde{\underset{\sim}{H}}(z)\}$ is bounded by a decaying exponential; more precisely, for any $\beta \in (\rho,1)$ there is a finite number b such that

$$|\underset{\sim}{H}_i| \leq b\beta^i \quad \forall i \in J_+ \qquad (10)$$

Theorem 4

Consider the multiple-input, multiple-output, linear, time-varying,

discrete system S_1 shown in Fig. 2. Let $\underset{\sim}{G}$ be the linear subsystem whose z-transfer function is given by

$$\underset{\sim}{\tilde{G}}(z) = \underset{\sim}{R}(1-z^{-1})^{-1} + \sum_{i=0}^{\infty} \underset{\sim}{G}_i z^{-i} \triangleq \underset{\sim}{R}(1-z^{-1})^{-1} + \underset{\sim}{\tilde{G}}_{\ell}(z) \qquad (11)$$

where $\underset{\sim}{R}$ is an nxn constant matrix and $\underset{\sim}{G}_{\ell} = \{\underset{\sim}{G}_i\}_0^{\infty} = \mathcal{Z}^{-1}\{\underset{\sim}{\tilde{G}}_{\ell}(z)\} \in \ell^1_{n\times n}$.
Let $\underset{\sim}{K} : J_+ \to R^{n\times n}$ be the memoryless, time-varying gain which is defined by

$$\underset{\sim}{K} \triangleq \underset{\sim}{\bar{K}} + \underset{\sim}{\hat{K}} = \underset{\sim}{\bar{K}} + \{\underset{\sim}{\ddot{K}}_i\}_0^{\infty} \qquad (12)$$

where $\underset{\sim}{\bar{K}}$ is an n$\times$n constant matrix and $\underset{\sim}{\hat{K}} \triangleq \{\underset{\sim}{\hat{K}}_i\}_0^{\infty} \in \ell^1_{n\times n}$.

Let $\det [\underset{\sim}{I} + (\underset{\sim}{R} + \underset{\sim}{G}_o)\underset{\sim}{\bar{K}} + (\underset{\sim}{R} + \underset{\sim}{G}_o)\underset{\sim}{\hat{K}}_i] \neq 0 \ \forall i \in J_+$ (13)
Under these conditions, if

$$\inf_{|z| \geq 1} |\det [\underset{\sim}{I} + \underset{\sim}{\tilde{G}}(z)\underset{\sim}{\bar{K}}]| > 0 \qquad (14)$$

and if either $\underset{\sim}{R} = \underset{\sim}{0}$ or $\underset{\sim}{R}\underset{\sim}{\bar{K}}$ is nonsingular, then

(a) $\underset{\sim}{u} \in \ell^{\infty}_n$ implies that $\underset{\sim}{w} \in \ell^{\infty}_n$

(b) $\underset{\sim}{u} \in \ell^1_n$ implies that $\underset{\sim}{w} \in \ell^1_n$, consequently $\lim_{m\to\infty} \underset{\sim}{w}_m = \underset{\sim}{0}$.

The same results also hold for $\underset{\sim}{e}$ and $\underset{\sim}{y}$.

The results presented above are important not only as far as linear systems are concerned but also because they constitute key tools for the study of nonlinear, time-invariant or time-varying systems. Here we give no proof for the above theorems. The full length paper will appear in the Journal of Franklin Institute (14).

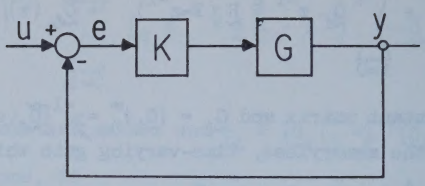


Fig. 1 Linear discrete feedback system $\underset{\sim}{S}$ under consideration. $\underset{\sim}{G}$ is a convolution-type linear time-invariant subsystem and $\underset{\sim}{K}$ is a (memoryless) time-varying gain. $\underset{\sim}{u}$ is the input and $\underset{\sim}{y}$ is the output.

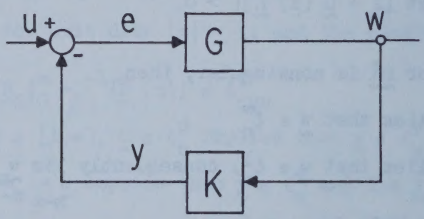


Fig. 2 System $\underset{\sim}{S}_1$ differs from system $\underset{\sim}{S}$ by the interchange in the order of the subsystems $\underset{\sim}{G}$ and $\underset{\sim}{K}$. $\underset{\sim}{u}$ is the input and $\underset{\sim}{w}$ is the output. $(\underset{\sim}{u},\underset{\sim}{y})$ is an input-output pair of $\underset{\sim}{S}$ if and only if $(\underset{\sim}{K}\underset{\sim}{u},\underset{\sim}{y})$ is an input-output pair of $\underset{\sim}{S}_1$.

Research sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016(Sup 6), and the Joint Services Electronics Program under Grant AFOSR-68-1488.

References

(1) a) E.I. Jury, "Sampled-Data Systems", New York: John Wiley and Sons, Inc., 1958.

b) E.I. Jury, "Theory and Application of the z-Transform Method", New York: John Wiley and Sons, Inc., 1964.

(2) H. Freeman, "Discrete-Time Systems", New York: John Wiley and Sons, Inc., 1965.

(3) B.C. Kuo, "Analysis and Synthesis of Sampled-Data Control Systems", Englewood Cliffs, N.J.: Prentice Hall, 1963.

(4) J.T. Tou, "Digital and Sampled-Data Control Systems", New YorkL McGraw-Hill, 1959.

(5) Y.Z. Tsypkin and E.I. Jury, "On the Theory of Discrete Systems", To appear in Automation and Remote Control.

(6) I.W. Sandberg, "On the Boundedness of Solutions of Nonlinear Integral Equations", Bell System Tech. J., Vol. 44, pp. 439-453, March, 1965.

(7) C.T. Chen, "L^p-Stability for Linear Time-Varying Feedback Systems", SIAM J. on Control, Vol. 6, pp. 186-193, 1966.

(8) C. Goffman and G. Pedrick, "First Course in Functional Analysis", Englewood Cliffs, N.J.: Prentice-Hall, 1965.

(9) N. Wiener, "The Fourier Integral and Certain of Its Applications", New York: Cambridge University Press, 1933.

(10) J. Dieudonné, "Foundations of Modern Analysis", New York: Academic Press, 1960.

(11) G.S. Jones, "Fundamental Inequalities for Discrete and Discontinuous Functional Equations", Jour. SIAM, Vol. 12, No. 1, pp. 43-57, March, 1964.

(12) W. Rudin, "Real and Complex Analysis", New York: McGraw-Hill, 1966.

(13) L.A. Zadeh and C.A. Desoer, "Linear System Theory - The State Space Approach", New York: McGraw-Hill, 1963.

(14) C.A. Desoer and M.Y. Wu, "Input-Output Properties of Multiple-Input Multiple-Output Discrete Systems: Part I", To appear in the Journal of Franklin Institute.

INPUT-OUTPUT PROPERTIES OF NONLINEAR DISCRETE SYSTEMS

M.Y. Wu
U. of Colorado, Boulder, Colo., formerly with U.C. Berkeley
C.A. Desoer
U.C. Berkeley

Abstract

This paper considers the input-output properties of nonlinear time-varying systems. Some recent results of linear discrete systems, the generalized forms of the small gain and passivity theorem and the multiplier are used to derive stability criteria for a very broad class of multiple-input, multiple-output nonlinear, time-varying discrete systems. The memoryless nonlinearities and the multipliers are not required to be noninteracting.

This paper considers the input-output properties of nonlinear time-varying systems. In this paper we consider the system model shown in Fig. 1.

The sequences $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$; $\underset{\sim}{e}_1$, $\underset{\sim}{e}_2$; $\underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ are in Σ^n. (Σ^n denotes the space of all sequences in R^n.) $\underset{\sim}{H}_1$, $\underset{\sim}{H}_2$: $\Sigma^n \to \Sigma^n$ are operators which can be linear or/and nonlinear, time-invariant or/and time-varying. The system $\underset{\sim}{\Sigma}$ is determinate, and described by the following system equations.

$$\underset{\sim}{e}_1 = \underset{\sim}{u}_1 - \underset{\sim}{y}_2 \tag{1}$$

$$\underset{\sim}{e}_2 = \underset{\sim}{u}_2 + \underset{\sim}{y}_1 \tag{2}$$

$$\underset{\sim}{y}_1 = \underset{\sim}{H}_1 \underset{\sim}{e}_1 \tag{3}$$

$$\underset{\sim}{y}_2 = \underset{\sim}{H}_2 \underset{\sim}{e}_2 \tag{4}$$

Some recent results of linear discrete systems, the generalized forms of the small gain and the passivity theorem and the multiplier are used to derive stability criteria for a very broad class of multiple-input, multiple-output, nonlinear, time-varying discrete systems. The main results are stated in the following three theorems:

Theorem 1A (Application of Small Gain Theorem)

Consider the system $\underset{\sim}{\Sigma}$ (Fig. 1) with $\underset{\sim}{H} = \underset{\sim}{G}$ being linear, time-invariant, nonanticipative subsystem and $\underset{\sim}{H}_2 = \underset{\sim}{\Phi}_t$ being time-varying memoryless nonlinearity. Let the input-output relation of the linear subsystem $\underset{\sim}{G}$ be defined in terms of its impulse response $\underset{\sim}{G}$ by the convolution

$$\underset{\sim}{y}_1 = \underset{\sim}{G} * \underset{\sim}{e}_1 \tag{16}$$

Let the open-loop z-transfer function of $\underset{\sim}{G}$ be of the form

$$\underset{\sim}{\tilde{G}}(z) = \underset{\sim}{R}\left(1 - z^{-1}\right)^{-1} + \sum_{i=0}^{\infty} \underset{\sim}{G}_i z^{-i} \triangleq \underset{\sim}{R}\left(1 - z^{-1}\right)^{-1} + \underset{\sim}{\tilde{G}}_\ell(z) \tag{16a}$$

where $\underset{\sim}{R}$ is an nXn constant matrix and $\underset{\sim}{G}_\ell \triangleq \{\underset{\sim}{G}_i\}_0^\infty = \mathcal{Z}^{-1}\{\underset{\sim}{\tilde{G}}_\ell(z)\} \epsilon \ell^1_{nXn}$. Let the time-varying memoryless nonlinearity $\underset{\sim}{\Phi}_t$ be described by a nonlinear function

$\underset{\sim}{\Psi}_t : \Sigma^n \times J_+ \to \Sigma^n$, ($J_+$ denotes the set of all nonnegative integers) which satisfies the condition that for some constant matrix $\underset{\sim}{K}$, some nonnegative number μ_2 and some constant ν_2

$$|\underset{\sim}{\Psi}_t(\sigma,m) - \underset{\sim}{K}\sigma| \leq \mu_2|\sigma| + \nu_2 \quad \forall \sigma \in \Sigma^n, \ \forall m \in J_+ \tag{17}$$

Under these conditions, if

$$\text{(a)} \qquad \operatorname{Inf}_{|z| \geq 1} |\det(\underset{\sim}{I} + \underset{\sim}{\tilde{G}}(z)\,\underset{\sim}{K})| > 0 \tag{18}$$

and if either $\underset{\sim}{R} = \underset{\sim}{0}$ or $\underset{\sim}{RK}$ is nonsingular,

$$\text{(b)} \qquad \|\underset{\sim}{H}\|_1 \mu_2 < 1 \tag{19}$$

where $\underset{\sim}{H} \triangleq \{\underset{\sim}{H}_i\}_0^\infty = \mathcal{Z}^{-1}\{[\underset{\sim}{I} + \underset{\sim}{\tilde{G}}(z)\,\underset{\sim}{K}]^{-1}\,\underset{\sim}{\tilde{G}}(z)\}$,

then for any fixed $p \in [1,\infty]$ $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$ in ℓ_n^p implies that $\underset{\sim}{e}_1$, $\underset{\sim}{e}_2$, $\underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ are also in ℓ_n^p.

Theorem 1B

Consider the system $\underset{\sim}{\mathcal{S}}$ (Fig. 1) with $\underset{\sim}{H}_1 = \underset{\sim}{G}$ being a linear, time-invariant, nonanticipative subsystem which is described by (16) and (16a) and $\underset{\sim}{H}_2 = \underset{\sim}{K}$ being a linear, time-varying gain $\underset{\sim}{K}$ which is specified by a sequence of nxn matrices $\{\underset{\sim}{K}_i\}_0^\infty$, where $|\underset{\sim}{K}_i| < \infty \ \forall i \in J_+$. Let the system $\underset{\sim}{\mathcal{S}}$ be determinate, i.e.,

$$\det[\underset{\sim}{I} + (\underset{\sim}{G}_o + \underset{\sim}{R})\,\underset{\sim}{K}_i] \neq 0 \ \forall i \in J_+$$

Under these conditions, if there is a constant matrix $\bar{\underset{\sim}{K}}$ such that $\underset{\sim}{K}_i \to \bar{\underset{\sim}{K}}$ as $i \to \infty$ and furthermore

$$\operatorname{Inf}_{|z| \geq 1} |\det[\underset{\sim}{I} + \underset{\sim}{\tilde{G}}(z)\,\bar{\underset{\sim}{K}}]| > 0 \tag{18a}$$

then for any fixed $p \in [1,\infty]$, $\underset{\sim}{u}_1$, $\underset{\sim}{u}_2$ in ℓ_n^p implies that $\underset{\sim}{e}_1$, $\underset{\sim}{e}_2$, $\underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ are also in ℓ_n^p.

Theorem 2A (Application of Passivity Theorem)

Consider the same system $\underset{\sim}{\mathcal{S}}$ as in Theorem 1A, where the linear time-invariant nonanticipative subsystem $\underset{\sim}{G}$ is described by (16) and (16a); the memoryless, time-varying nonlinearity $\underset{\sim}{\Phi}_t$ is described by nonlinear function $\underset{\sim}{\Psi}_t : \Sigma^n \times J_+ \to \Sigma^n$ which has the following properties:

N1. for some constant nxn matrix $\underset{\sim}{K}$

$$(\underset{\sim}{\sigma}_1 - \underset{\sim}{\sigma}_2)' [\underset{\sim}{\psi}_t(\underset{\sim}{\sigma}_1, m) - \underset{\sim}{\psi}_t(\underset{\sim}{\sigma}_2, m)] \geq (\underset{\sim}{\sigma}_1 - \underset{\sim}{\sigma}_2)' \underset{\sim}{K}(\underset{\sim}{\sigma}_1 - \underset{\sim}{\sigma}_2)$$

$$\forall\, \underset{\sim}{\sigma}_1, \underset{\sim}{\sigma}_2 \in \Sigma^n, \forall\, m \in J_+, \tag{20}$$

N2. $$\underset{\sim}{\psi}_t(-\underset{\sim}{\sigma}, m) = -\underset{\sim}{\psi}_t(\underset{\sim}{\sigma}, m) \quad \forall\, \underset{\sim}{\sigma} \in \Sigma^n, \forall\, m \in J_+ \tag{21}$$

Let $\underset{\sim}{M}$ be the multiplier whose z-transform function is of the form

$$\underset{\sim}{\tilde{M}}(z) = \sum_{i=0}^{\infty} \underset{\sim}{M}_i\, z^{-1} \tag{22}$$

and satisfies the following conditions:

M1. $$\underset{\sim}{M} \overset{\Delta}{=} \{\underset{\sim}{M}_i\}_0^\infty = \mathcal{Z}^{-1}\{\underset{\sim}{\tilde{M}}(z)\} \in \ell^1_{nxn}$$

M2, $$\inf_{|z| \geq 1} |\det \underset{\sim}{\tilde{M}}(z)| > 0 \tag{23}$$

M3. for all $i \in J_+$, all elements of $\underset{\sim}{M}_i$ are such that

$$(m_i)_{\alpha\alpha} \geq \sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{n} |(m_i)_{\alpha\beta}| \text{ and } (m_i)_{\beta\beta} \geq \sum_{\substack{\alpha=1 \\ \alpha\neq\beta}}^{n} |(m_i)_{\alpha\beta}| \tag{24}$$

Under these conditions, if

(i) for the constant matrix $\underset{\sim}{K}$ defined in (20)

$$\inf_{|z| \geq 1} |\det [\underset{\sim}{I} + \underset{\sim}{G}(z)\underset{\sim}{K}]| > 0 \tag{25}$$

and if $\underset{\sim}{R} = \underset{\sim}{0}$ or $\underset{\sim}{RK}$ is nonsingular,

(ii) for some number $\delta > 0$,

$$\inf_{|z| = 1} \lambda\, \{\underset{\sim}{\tilde{M}}(z)[\underset{\sim}{I} + \underset{\sim}{\tilde{G}}(z)\underset{\sim}{K}]^{-1}\underset{\sim}{\tilde{G}}(z) + \underset{\sim}{\hat{G}}'(\bar{z})[\underset{\sim}{I} + \underset{\sim}{K}'\underset{\sim}{\tilde{G}}'(\bar{z})]^{-1}\underset{\sim}{\tilde{M}}'(\bar{z})\} \geq \delta > 0 \tag{26}$$

where $\lambda\,\{\underset{\sim}{W}\}$ denotes the least eigenvalue of the matrix $\underset{\sim}{W}$, then for all $\underset{\sim}{u}_1, \underset{\sim}{u}_2$ in ℓ^2_n, $\underset{\sim}{e}_1, \underset{\sim}{e}_2, \underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ are in ℓ^2_n.

Note that in the results presented above the memoryless nonlinearities and the multipliers are not required to be noninteracting.
The above results are stated without proofs. The full length paper will appear in the J. of the Franklin Institute [20].

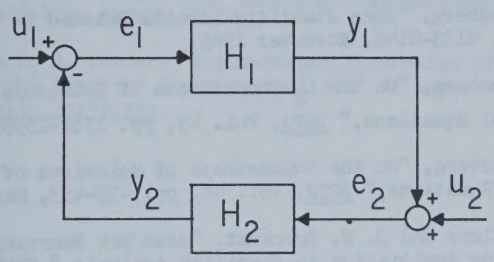


Fig. 1. The system $\mathcal{S}$ under consideration.

Research sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016(Sup 6), and the Joint Services Electronics Program under Grant AFOSR-68-1488.

References

[1] G. Zames, "On the Input-Output Stability of Time-Varying Nonlinear Feedback Systems," Pt. I, IEEE Trans., Vol AC-11, No. 2, pp. 228-238, April 1966.

[2] E. S. Kuh and R. A. Rohrer, Theory of Linear Active Networks, San Francisco: Holden-Day, 1967.

[3] I. W. Sandberg, "Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations," BSTJ, Vol. 44, pp. 871-898, May-June 1965.

[4] I. W. Sandberg, "Some Stability Results Related to Those of V.M.Popov," BSTJ, pp. 2133-2148, November 1965.

[5] I. W. Sandberg, "On the L_2-Boundedness of Solutions of Nonlinear Functional Equations," BSTJ, Vol. 43, pp. 1581-1599, July 1964.

[6] I. W. Sandberg, "On the Boundedness of Solutions of Nonlinear Integral Equations," BSTJ, Vol. 44, pp. 439-453, March 1965.

[7] J. C. Willems and R. W. Brockett, "Some New Rearrangement Inequalities Having Application in Stability Analysis," IEEE Trans., Vol. AC-13, No. 5, pp. 539-549, October 1968.

[8] R. P. O'Shea and M. I. Younis, "A Frequency-Time Domain Stability Criterion for Sampled-Data Systems," IEEE Trans., Vol. AC-12, No. 6, pp. 719-724, December 1967.

[9] Y. Z. Tsypkin, "A Criterion for Absolute Stability of Automatic Pulse Systems with Monotonic Characteristics of the Nonlinear Element," Dokl. Akad. Nauk (USSR), Vol. 155, pp. 1029-1032, April 1964.

[10] E. I. Jury and B. W. Lee, "On the Stability of a Certain Class of Nonlinear Sampled-Data Systems," IEEE Trans., Vol. AC-9, pp. 51-61, January 1964.

[11] E. I. Jury and B. W. Lee, "A Stability Theory for Multi-Nonlinear Control Systems," Proc. 3rd IFAC Conference, London, June 1966, paper 28A.

[12] C. T. Chen, "On the Stability of Nonlinear Sampled-Data Feedback Systems," J. Franklin Inst., Vol. 280, pp. 316-324, October 1965.

[13] C. A. Desoer and M. Y. Wu, "Input-Output Properties of Multiple-Input, Multiple-Output Discrete Systems, Part I," to appear in J. of the Franklin Institute.

[14] Y. S. Cho and K. S. Narendra, "Stability of Nonlinear Time-Varying Feedback Systems," Proc. 5th Allerton Conference, October 1967, pp. 249-258.

[15] T. E. Stern, "The Network Analog as a Heuristic Tool for Generating Stability Criteria," Proc. 4th Allerton Conference, October 1966, pp. 819-826.

[16] M. Y. Wu, "Stability of Nonlinear Feedback Systems," Ph.D. Dissertation, University of California, Berkeley, July 1968.

[17] J. C. Willems, M. Gruber and R. P. O'Shea, "Comments on a Combined Frequency-Time Domain Stability Criterion for Autonomous Continuous Systems," IEEE Trans., Vol. AC-12, pp. 217-219, April 1967.

[18] P. L. Falb and G. Zames, "On Cross-Correlation Bounds and the Positivity of Certain Nonlinear Operators," IEEE Trans., Vol. AC-12, pp. 219-221, April 1967.

[19] C. T. Chen, "Bounded-Input, Bounded-Output Stability of Linear Time-Varying Feedback Systems," J. of the Franklin Inst., Vol. 286, No. 2, pp. 128-134, August 1968.

[20] M. Y. Wu and C. A. Desoer, "Input-Output Properties of Multiple-Input Multiple-Output Discrete Systems: Part II," to appear in the Journal of the Franklin Institute.

SOME RESULTS IN PASSIVITY AND STABILITY OF SYSTEMS WITH A STATE REPRESENTATION

R. F. ESTRADA
Bell Telephone Laboratories
Whippany, New Jersey

C. A. DESOER
University of California
Berkeley, California

ABSTRACT

The concept of passivity is developed for systems with a state representation (dynamical systems). This development of passivity is shown to be (1) related to operator passivity as defined by Zames, Falb, Wu and others, and (2) intuitively related to network passivity. Passivity as defined in the paper leads to stability theorems for feedback systems. New results are obtained (1) in area of bounded-input bounded-state stability and (2) for general time-varying feedback systems.

INTRODUCTION

The intent of this paper is to give a general view of stability results that can be obtained from passivity concepts. In the first part stability theorems for general feedback systems are developed based on passivity as defined in the paper. In the second part, applications of these theorems to more specific systems are considered. The proofs of the theorems and lemmas given here can be found in [1] and [2] which offer a more comprehensive view of the material.

GENERAL THEORY

In this paper we will be dealing with systems with a state representation or dynamical systems. These systems are discussed in [3] and [4] as well as in many other places. The particular definition of a dynamical system that will be used here is the following.

Definition 1. A System S is said to be a dynamical system it if consists of seven objects

$$S = (U,\mathcal{U},X,Y,\mathcal{Y},F_o,F_s) \tag{1}$$

which have several properties. The seven objects are:

U is a linear space called the space of input values. [U is often R or R^n].

$\mathcal{U}$ is a function space: its element map $R_+ \to U$ where $R_+ = [0,\infty)$. An element of $\mathcal{U}$ is called an input. For $t \in R_+$, $u(t) \in U$ is called the value of u at time t.

X is a linear space called the state space and if $x \in X$ then x is called a state.

Y is a linear space called the space of output values. [Y is often R or R^n].

$\mathcal{Y}$ is a function space: its element map $R_+ \to Y$. An element of $\mathcal{Y}$ is called an output. For $t \in R_+$ $y(t) \in Y$ is called the value of y at time t.

F_o and F_s are maps

$$F_o : R_+ \times R_+ \times X \times \mathcal{U} \to Y \qquad F_s : R_+ \times R_+ \times X \times \mathcal{U} \to X$$

such that

$$F_o\left(t,t_0,x_{t_0},u\right) = y(t) \qquad F_s\left(t,t_0,x_{t_0},u\right) = x_t \tag{2}$$

and in addition

$$\left.\begin{aligned} F_o\left(\cdot,t_0,x_{t_0},u\right) &\in \mathcal{Y} \\ F_s\left(t_0,t_0,x_{t_0},u\right) &= x_t \end{aligned}\right\} \quad \text{for any} \quad \left(t_0,x_{t_0},u\right) \in R_+ \times X \times \mathcal{U}$$

These maps must have the following two properties:

Nonanticipation Property

For all $t_0,t_1 \in R_+$; for all $x_{t_0} \in X$; for all $\tilde{u},u \in \mathcal{U}$ such that $u(t') = \tilde{u}(t')$ for all $t' \in [t_0,t_1]$,

$$\left.\begin{aligned} F_o\left(t,t_0,x_{t_0},u\right) &= F_o\left(t,t_0,x_{t_0},\tilde{u}\right) \\ F_s\left(t,t_0,x_{t_0},u\right) &= F_s\left(t,t_0,x_{t_0},\tilde{u}\right) \end{aligned}\right\} \quad \text{for all } t \in [t_0,t_1]$$

Separation Property or Semi-group Axiom

For all $t_0,t_1,t_2 \in R_+$ such that $t_0 \le t_1 \le t_2$; for all $x_{t_0} \in X$; for all $u \in \mathcal{U}$,

$$F_s\left[t_2,t_1,F_s\left(t_1,t_0,x_{t_0},u\right),u\right] = F_s\left(t_2,t_0,x_{t_0},u\right)$$

$$F_o\left[t_2,t_1,F_s\left(t_1,t_0,x_{t_0},u\right),u\right] = F_o\left(t_2,t_0,x_{t_0},u\right)$$

In addition, the following assumption is assumed to hold throughout.

Assumption 1. (i) X is normed with norm $|\cdot|$, $Y = U$, and has an inner product $(\cdot|\cdot)$ and induced norm $|u(t)|_1^2 = \left(u(t)|u(t)\right)$. (ii) $\mathcal{U} = \mathcal{Y} = L_{2e}^Y$ where $u \in L_{2e}^Y$ iff $\int_0^T |u(t)|_1^2 dt < \infty$ for all finite T

Notation: The space L_2^Y, naturally, is defined by

$$u \in L_2^Y \quad \text{iff} \quad \int_0^\infty |u(t)|_1^2 dt < \infty$$

Note: The notation x_t rather than $x(t)$ will be used for the state at time t when it is not certain that X is of finite dimension. The important sets defined in the following definitions will be discussed in an intuitive manner shortly.

Definition 2. Let S be a dynamical system meeting Assumption 1. Then, associated with S is the set A_S defined by: Every member of the set A_S consists of a pair of continuous functions (E,P) called the pseudo-energy available function and the pseudo-power dissipation function and $(E,P) \in A_S$ iff $E:X\times R_+ \to R$, $P:Y\times U\times X\times R_+ \to R$, and are such that for every $\left(x_{t_0},t_0\right) \in X\times R_+$

$$E\left(x_{t_0},t_0\right) = \sup_{e\in L_2^Y} \left\{ \begin{array}{l} -\int_{t_0}^{\infty} \left(F_o\left(t,t_0,x_{t_0},e\right)\middle|e(t)\right)dt \\ \\ +\int_{t_0}^{\infty} P\left[F_o\left(t,t_0,x_{t_0},e\right),e(t),F_s\left(t,t_0,x_{t_0},e\right),t\right]dt \end{array} \right. \tag{3}$$

Note: An alternate way of defining the set A_S would be: Given a continuous P function, if, for any $\left(x_{t_0},t_0\right) \in X\times R_+$, the supremum is finite in Equation (3) and we call that number $E\left(x_{t_0},t_0\right)$ then, if E is continuous on $X\times R_+$, $(E,P) \in A_S$. Also, if $(E,P) \in A_S$ and P is non-negative then E is non-negative. This is true since $e = 0 \in L_2^Y$.

Definition 3. For a dynamical system S and a continuous function $E:X\times R_+ \to R$, E' (which we will call the generalized time derivative of E along the trajectories of S) is defined by

$$E'(x_t,u,t) \triangleq \lim_{h\to 0^+} \sup \frac{1}{h}\left[E\left(F_s(t+h,t,x_t,u),t+h\right) - E(x_t,t)\right] \tag{4}$$

The set defined in the following definition is closely related to the set A_S.

Definition 4. For a dynamical system S meeting Assumption 1, the set B_S is defined by: Every member of the set B_S is a pair of continuous functions (E,P) and $(E,P) \in B_S$ iff $E:X\times R_+ \to R$, $P:Y\times U\times X\times R_+ \to R$, and are such that for every $(x_t,t,u) \in X\times R_+\times L_{2e}^U$

$$E'(x_t,u,t) \le \left(F_o(t,t,x_t,u)\middle|u(t)\right) - P\left(F_o(t,t,x_t,u),u(t),x_t,t\right) \tag{5}$$

In order to get some intuitive feeling about the sets A_S and B_S, let us digress and consider the network N shown in Figure 1. Here u is the voltage input and y is the current output. In addition, it is assumed that the structure of N is completely known and that:

(1) An acceptable state for N, x, has been found.

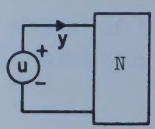


Fig. 1

(2) Functions E_s and P_d have been found for N such that $E_s\left(x(t),t\right)$ is the energy stored in N (through capacitors, etc.) and $P_d\left(y(t),u(t),x(t),t\right)$ is the power being dissipated by N. (Negative values of P_d imply that power is being generated in N.)

(3) The equations of N have been solved and functions F_o and F_s have been found such that $y(t) = F_o\left(t,t_0,x(t_0),u\right)$ and $x(t) = F_s\left(t,t_0,x(t_0),u\right)$.

Note that $y(t)u(t)$ represents power entering N and the passivity of N is loosely connected to how positive the function P_d is. The conservation of energy, applied to N, provides that

$$E_s\left(x(t_0),t_0\right) - E_s\left(x(t_f),t_f\right) = \int_{t_0}^{t_f}\Bigg[-F_o\left(t,t_0,x(t_0),u\right)u(t) + P_d\Big[F_o\left(t,t_0,x(t_0),u\right),u(t),F_s\left(t,t_0,x(t_0),u\right),t\Big]\Bigg]dt \tag{6}$$

Assuming that some possible inputs are square integrable (L_2), if $t_f \to \infty$ and the supremum of both sides are taken with respect to $u \in L_2$, then

$$E_s\left(x(t_0),t_0\right) - \inf_{u\in L_2} E_s\left(x(\infty),\infty\right) = \sup_{u\in L_2}\int_{t_0}^{\infty}\Bigg[-F_o\left(t,t_0,x(t_0),u\right)u(t) + P_d\Big[F_o\left(t,t_0,x(t_0),u\right),u(t),F_s\left(t,t_0,x(t_0),u\right),t\Big]\Bigg]dt \tag{7}$$

Now $\inf_{u\in L_2} E_s\left(x(\infty),\infty\right) \geq 0$ and is only a function of $\left(x(t_0),t_0\right)$; therefore we have defined a function $E_a\left(x(t_0),t_0\right)$[energy available] such that

$$E_a\left(x(t_0),t_0\right) = \sup_{u\in L_2}\int_{t_0}^{\infty}\Big[-F_o\left(t,t_0,x(t_0),u\right)u(t) + P_d\left(F_o\left(t,t_0,x(t_0),u\right),u(t),E_s\left(t,t_0,x_{t_0},u\right),t\right)dt\Big] \tag{8}$$

and

$$E_a\left(x(t_0),t_0\right) \leq E_s\left(x(t_0),t_0\right)$$

The conservation of power applied to N provides that

$$\frac{dE_s\left(x(t),t\right)}{dt} = F_o\left(t,t,x(t),u\right)u(t) - P_d\left(F_o\left(t,t,x(t),u\right),u(t),x(t),t\right) \qquad (9)$$

where $\frac{dE_s\ x(t),t}{dt}$ is the total time derivative of E_s along the state trajectories (in the same sense as Definition 3).

At this point the relationship between the sets A_S nnd B_S and the network N is clear. If E_{aN}, E_{sN} and P_{dN} are the energy available, energy stored and power dissipation function for N and if S is the mathematical description of N, then, comparing Equations 3 and 8 $(E_{aN},P_{dN}) \in A_S$ and, comparing Equations 5 and 9 $(E_s,P_{dN}) \in B_S$. However, S could be the description of many different networks, say N_i (of the same form as N), $i = 1,2, \ldots$ then $\left(E_{aN_i},P_{dN_i}\right) \in A_S$, $\left(E_{sN_i},P_{dN_i}\right) \in B_S$ $i = 1,2, \ldots$. There also may be members of A_S and B_S for which there is no corresponding network. We have seen that, roughly speaking, the set B_S is the differential form of A_S. In fact the following lemma can be proved [1], [2].

Lemma 1. For a dynamical system S, if $(E,P) \in A_S$ then $(E,P) \in B_S$.

Lemma 2. If $(E,P) \in A_S$ or $(E,P) \in B_S$, then for all $h > 0$, $(x_t,t,u) \in X\times R_+\times L^U_{2e}$

$$E\left(F_s(t+h,t,x_t,u),t+h\right) - E(x_t,t)$$

$$\leq \begin{cases} \displaystyle\int_t^{t+h} \left(F_o(t',t,x_t,u|u(t'))\right)dt' \\ \displaystyle\int_t^{t+h} -P\left(F_o(t',t,x_t,u),u(t'),F_s(t',t,x_t,u),t'\right)dt' \end{cases} \qquad (10)$$

That $(E,P) \in A_S$ implies Inequality (10) becomes evident during the proof of Lemma 1. That $(E,P) \in B_S$ implies Inequality (10) is proved in [5] and [6].

For the network N, it was stated that the degree of passivity of N was loosely connected to positiveness of P_d. We will carry this idea over to dynamical systems and define passivity for them.

Definition 4. A dynamical system S, meeting Assumption 1, is said to be

(i) passive if $(E,0) \in A_S$ where E is any continuous function such that $E:X\times R_+ \to R$

(ii) strictly passive if there exist $(E,P) \in A_S$ and $\delta > 0$ such that $P \geq \delta|u(t)|_1^2$

(iii) very strictly passive is there exist $(E,P) \in A_S$,

$$\delta_1 > 0 \text{ and } \delta_2 > 0 \text{ such that } P \geq \delta_1 |u(t)|_1^2 + \delta_2 |x_t|^2.$$

Note that $P \geq \delta_1 |u(t)|_1^2 + \delta_2 |x_t|^2$ stands for $P\left(F_o(t,t,x_t,u),u(t),x_t,t\right) \geq \delta_1 |u(t)|_1^2 + \delta_2 |x_t|^2$ for all $(x_t,t,u) \in X \times R_+ \times L_{2e}^U$. Passivity in the "differential sense" could be defined by using B_S rather than A_S. In this case the conditions for passivity would be very similar to those of Rohrer's [7], for the class of systems he considered. If one considers $\left(x_{t_0},t_0\right)$ fixed then $F_o\left(\cdot,t_0,x_{t_0},u\right)$ is an operator mapping L_{2e}^U into L_{2e}^U. Thus if S is passive in the sense of Definition 4,

$$-E\left(x_{t_0},t_0\right) \leq \int_{t_0}^{\infty} \left(F_o\left(t',t_0,x_{t_0},u\right)|u(t')\right)dt' \text{ for all } u \in L_{2e}^U$$

and the operator F_o is passive in the extended sense defined by Wu [8].

We will see shortly that the different degrees of passivity for a dynamical system are associated with different types of stability.

The main application, of the concepts previously defined, to stability is for the system S_c shown in Figure 2.

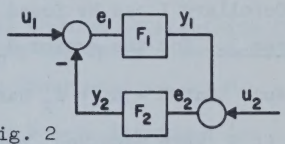


Fig. 2

In S_c, F_1 and F_2 are both dynamical systems meeting Assumption 1, that is

$$F_1 = \left\{E_1,\mathcal{E}_1,X_1,Y_1,\mathcal{Y}_1,F_{10},F_{1s}\right\}, \quad F_2 = \left\{E_2,\mathcal{E}_2,X_2,Y_2,\mathcal{Y}_2,F_{20},F_{2s}\right\}$$

and it must be that $E_1 = Y_1 = U_1 = E_2 = Y_2 = U_2 = H$. [H is an inner product space with inner product $(\cdot|\cdot)$.] We assume that the interconnection in determinate [4] and that $X \triangleq X_1 \times X_2$ can be taken to be the state space for S_c. Let $|x|^2 = |(x_1,x_2)|^2 = |x_1|_{X_1}^2 + |x_2|_{X_2}^2$. Let $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 = L_{2e}^{H \times H}$ and we assume $\mathcal{Y}$ can be taken to be $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 = L_{2e}^{H \times H}$. Thus the system S_c is a dynamical system meeting Assumption 1.

The following two theorems show that it is not necessary to know the F_s and F_o maps for S_c explicitly in order to find members of the sets A_{S_c} and B_{S_c}.

Theorem 1. For the system S_c, if $(E_1,P_1) \in A_{F_1}$ and $(E_2,P_2) \in A_{F_2}$, then $(E_1+E_2,P_1+P_2) \in A_{S_c}$.

Theorem 2. For the system S_c, if $(E_1,P_1) \in B_{F_1}$ and $(E_2,P_2) \in B_{F_2}$ then $(E_1+E_2,P_1+P_2) \in B_{S_c}$.

The proofs of these theorems are based on the identity $\left(y_1(t)|u_1(t)\right) + \left(y_2(t)|u_2(t)\right) = \left(y_1(t)|e_1(t)\right) + \left(y_2(t)|e_2(t)\right)$ and can be found in [1] and [2].

Using Lemmas 1 and 2 and Theorems 1 and 2, many stability theorems for the system S_c can be proved. We will see that sufficient conditions for: Liapunov stability will be associated with passivity, L_2-input L_2-output stability with strict passivity, and bounded-input bounded-state stability with very strict passivity. Since we will be dealing with Liapunov stability we will need the definition of a positive definite function.

Definition 5. A function $E:X\times R_+ \to R$ is said to have property G if it is continuous on $X\times R_+$ and if there exist two functions $\alpha:R_+ \to R_+$, $\beta:R_+ \to R_+$ which are continuous, monotonically increasing, such that $\alpha(0) = 0$, $\beta(s) > 0$, $s > 0$ and, $\beta(s) \to \infty$ as $s \to \infty$, and for which $\alpha(|x|) \leq E(x,t) \geq \beta(|x|)$ for all $(x,t) \in X\times R_+$.

Notation. Let $|u|_\infty \triangleq \sup_{t\in R_+} |u(t)|_1$. The proofs of Theorems 3 through 5 and Corollary 1 can be found in [1] and [2].

Theorem 3. For the system S_c, if there exist $(E_1,P_1) \in A_{F_1}$ and $(E_2,P_2) \in A_{F_2}$ such that $E = E_1 + E_2$ has property G and $P_1 + P_2 \geq 0$ (this would be true if S_c were passive), then S_c is uniformly stable; that is, given any $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $t_0 \in R_+$, $|x_{t_0}| \leq \delta$ implies $\left|F_s\left(t,t_0,x_{t_0},0\right)\right| \leq \varepsilon$ for all $t \geq t_0$.

This theorem is easily proved when it is recognized that E is a "Liapunov function".

Theorem 4. For the system S_c, if there exist $(E_1,P_1) \in A_{F_1}$ and $(E_2,P_2) \in A_{F_2}$ and $\delta > 0$ such that

$$P_1 + P_2 \geq \delta\left(|e_1(t)|_1^2 + |e_2(t)|_1^2\right)$$

or in particular if F_1 and F_2 are both strictly passive, then for any $\left(x_{t_0},t_0\right) \in X\times R_+$ if $u_1,u_2 \in L_2^H$ then $y_1,y_2 \in L_2^H$.

Theorem 4 is an example of the many L_2-input, L_2-output theorems based on passivity. Further examples based on operator passivity can be found in [8], [11], and [12]. Examples based on passivity (Definition 4) are found in [2]. Theorem 4 is proved by applying Lemma 2 to the (E_1+E_2,P_1+P_2) pair found for S_c. The results of Lemma 2 give a form characteristic of L_2-input L_2-output theorems based on operator passivity.

Theorem 5. For the system S_c, if there exist $(E_1,P_1) \in A_{F_1}$; $(E_2,P_2) \in A_{F_2}$; $\delta_a > 0$; δ_b,δ_c with $\delta_b+\delta_c > 0$; δ_d,δ_e with $\delta_d+\delta_e > 0$; such that

$$P_1+P_2 \geq \begin{cases} \delta_a|x_t|^2 + \delta_b|y_1(t)|_1^2 + \delta_c|y_1(t)+u_2(t)|_1^2 \\ + \delta_d|y_2(t)|_1^2 + \delta_e|u_1(t)-y_2(t)|_1^2 \end{cases}$$

and $E \triangleq E_1 + E_2$ has property G, then S_c is bounded-input bounded state stable (BIBS stable) i.e. for any $x_{t_0} \in X$, $u \in \mathcal{U}$ with $|u|_\infty < \infty$ there exists $N\left(|x_{t_0}|, |u|_\infty\right)$ such that, for all $t_0 \in R_+$, $|F_s\left(t,t_0,x_{t_0},u\right)| \leq N$ for all $t \geq t_0$.

The proof of Theorem 5 is based on the extension of the work of Variaya and Liu [9] to dynamical systems which is done in [10].

Corollary 1. The results of Theorems 3 and 5 remain valid if A_{F_1} and A_{F_2} is replaced by B_{F_1} and B_{F_2}. If, in addition in theorem 4, $E_1 + E_2$ is non-negative then its results remain valid when A → B.

APPLICATIONS

In the General Theory, it was determined that problem of stability for a feedback system S_c could be approached by finding appropriate members of the sets A_{F_1} and A_{F_2} $\left(B_{F_1}$ and $B_{F_2}\right)$ which fit into Theorems 3, 4, or 5 (Corollary 1). In [1] and [2] members of the A or B sets are found for several typical systems (open loop). Feedback systems are then considered with these typical systems as subsystems (F_1 or F_2). We will present here some of the results of this procedure for two feedback systems.

System S_c^1. Consider the system S_c^1 shown in Figure 3. In S_c^1, H is a linear time-invariant system determined by its impulse response h. Its output map (F_o) is given by

$$y_1(t) = x_{1t_0}(t-t_0) + \int_{t_0}^{t} h(t-\tau)e_1(\tau)d\tau$$

where $x_{1t_0}(\cdot):R_+ \to R$ is the state of H at t_0. The state space for H, X_h, is defined by $x \in X_h$ iff there exists $\bar{u}:(-\infty,0) \to R$ with $\bar{u} \in L_2(-\infty,0)$ and $x(t) = \int_{-\infty}^{0} h(t-\tau)\bar{u}(\tau)d\tau$ for all $t \geq 0$. X_h has the L_2 inner product denoted $\langle \cdot | \cdot \rangle$. The state map of $H(F_s)$ is given by

$$F_s\left(t,t_0,x_{t_0},u\right) = \begin{cases} x_t & \text{for } t \geq t_0 \\ 0 & \text{for } t < t_0 \end{cases}$$

where

$$x_t:R_+ \to R:t_1 \mapsto x_{t_0}(t_1+t-t_0) + \int_{t_0}^{t} h(t+t_1-\tau)u(\tau)d\tau \qquad (11)$$

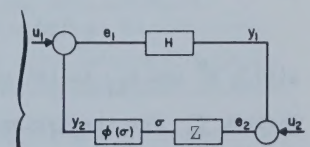


Fig. 3

The definition of the F_s map should be interpreted as: the state at time t, x_t, is a function defined on R_+. The value of the function x_t at $t_1 \in R_+$ is

$$x_{t_0}(t_1+t-t_0) + \int_{t_0}^{t} h(t+T_1-\tau)u(\tau)d\tau$$

It can be easily shown that H is a dynamical system. Also we will denote $\int_0^\infty tx^2(t)dt = \langle x|Tx\rangle$ for $x \in X_h$. $\phi(\sigma)$ is a memoryless nonlinear amplifier obeying a local Lipschitz condition and Z is the system governed by the equation $q\dot{\sigma}(t) + p\sigma(t) = u_2(t)$.

Theorem 6. For the system S_c^1, if there exist $k \in R$, $M \in R$ number $\gamma > 0$ and a number $\varepsilon > 0$ such that

(hi) $\mathrm{Re}\left[H(j\omega) + 1/k\right] \geq \gamma > 0$ for all $\omega \in R$

(hii) $h(t) \in L_1 \cap L_2$ and $th(t) \in L_1 \cap L_2$

(hiii) $\langle x|Tx\rangle \leq M\langle x|x\rangle$ for all $x \in X_h$

(ni) $\varepsilon\sigma^2 \leq \sigma\phi(\sigma) \leq (pk-\varepsilon)\sigma^2$ for all $\sigma \in R$

then there exists a positive definite bounded linear operator C, $\delta_a > 0$, $\delta_b > 0$ and $(E,P) \in A_{S_c^1}$ such that

$$E(x_t,t) = \langle x_{lt}|(C+\delta_a T+\delta_b)x_{lt}\rangle + q\int_0^{\sigma(t)} \phi(\sigma)d\sigma$$

E has property G (Definition 5) and

$$P \geq \left(-\frac{1}{k} + \frac{\gamma}{2}\right)e_1^2(t) + \delta_a|x_{lt}|^2 + \delta_b y_1^2(t) + \frac{1}{k}y_2^2(t) + \varepsilon(pk-\varepsilon)\sigma^2(t)$$

Thus by Theorem 5, S_c^3 is BIBS stable.

System S_c^2. Consider the system S_c^2 shown in Figure 4. In S_c^2, S^{tv} is the system governed by the equations $\dot{x}(t) = A(t)x(t) + b(t)e_1(t)$

$$y_1(t) = c(t)'x(t) + d(t)e_1(t)$$

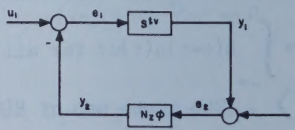


Fig. 4

$x(t) \in R^n$ and $N_{Z\phi}$ is the return loop system of S_c^1.

Theorem 7. For the system S_c^2, if there exists $\delta_a \in R$, $\varepsilon > 0$ and a positive semi-definite symmetric n×n matrix K(t) such that

(a) $d(t) - \delta_a > 0$ for all $t \in R_+$

(b) $K(t)$ satisfies

$$\dot{K}(t) + K(t)A(t) + A(T)'K(t) + \left[K(t)b(t) - c(t)\right]\left[d(t) - \delta_a\right]^{-1}\left[K(t)c(t) - b(t)\right]' = 0 \quad \text{for all } t \in R_+$$

(c) $\varepsilon\sigma^2 \leq \sigma\phi(\sigma) \leq \left(\frac{p}{\delta_a} - \varepsilon\right)\sigma^2$ for all $\sigma \in R$

then there exists $(E,P) \in B_{S_c^2}$ such that

$$E(x_t,t) = \frac{1}{2}x_1(t)'K(t)x_1(t) + q\int_0^{\sigma(t)} \phi(\sigma',t)d\sigma'$$

E is non-negative, and

$$P \geq \delta_a e_1^2(t) + \left(\frac{p}{k - \frac{\varepsilon}{2}}\right) y_2^2(t)$$

If, in addition $u_1 = 0$ and $u_2 \in L_2$, then, $y_2 \in L_2$.

REFERENCES

[1] Estrada, R. F. and C. A. Desoer, "Passivity and Stability of Systems with a State Representation," to be published.

[2] Estrada, R. F., Input Output Stability in Nonlinear Systems Using Lyapunov Methods, Ph.D. Dissertation, Univ. of Calif., Berkeley, 1969.

[3] Kalman, R. E., P. L. Falb and M. A. Arbib, Topics in Mathematical Systems Theory, New York: McGraw-Hill, 1969.

[4] Zadeh, L. A. and C. A. Desoer, Linear System Theory, New York: Academic Press, 1964.

[5] LaSalle, J. P., "Stability Theory for Ordinary Differential Equations," Journal of Differential Equations, Vol. 4, No. 1, January 1968.

[6] McShane, E. J., Integration, Princeton, N. J.: Princeton Univ. 1944.

[7] Rohrer, R. A., "Lumped Network Passivity Criteria," IEEE Trans. on Circuit Theory, Vol. CT-15, March 1968.

[8] Wu, M., Stability of Nonlinear Feedback Systems, Ph.D. Dissertation, University of California, Berkeley, 1968.

[9] Variaya, P. P. and R. Liu, "Bounded-Input Bounded-Output Stability of Nonlinear Time-Varying Differential Systems," Siam Journal Control, Vol. 4, No. 4, pp. 648-704, 1966.

[10] Estrada, R. F. and C. A. Desoer, "Bounded-Input Bounded-State Stability of Systems with a State Representation," to be published.

[11] Zames, G., "On the Input-Output Stability of Time Varying Nonlinear Feedback Systems, Parts I and II," IEEE Trans. on Automatic Control, Vol. AC-11, April and July 1966.

[12] Zames, G. and P. Falb, "On the Stability of Systems with Monotone and Odd Monotone Nonlinearities," IEEE Trans. on Automatic Control, AC-12, April 1967.

[13] Baker, R. and A. Bergen, "Lyapunov Stability and Lyapunov Functions of Infinite Dimensional Systems," Joint Automatic Control Conf., Ann Arbor, Michigan 1968.

A CLASS OF SYSTEM ESTIMATORS FOR ROUTING PROBLEMS

MARLIN H. MICKLE, RONALD G. HOELZEMAN, JAMES T. CAIN
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

ABSTRACT

The solution of large scale sequential decision problems such as the minimum route problem using heuristic techniques in general involves some form of estimator to predict system behavior. This paper reports the results of studying four estimators for a particular class of systems. Through the procedure which is established, a basis of comparison and method of evaluation of other estimators is obtained.

INTRODUCTION

Recent interest in routing problems and sequential decision problems has brought about the need for studying various methods of solution. The magnitude of many physical systems is such that classical methods of solution are too slow or otherwise not practical for large scale system modeling. The use of heuristic techniques appears to be one possible solution. A recent work in this area[1] has developed a formal basis for such determinations. The proposed algorithm involves an evaluation function similar to the policy equation used in another algorithm for modeling such systems[2]. In reference 1, the function, denoted by h(i), is an estimator for the unknown path portion of the result of the impending decision.

The estimator, in general, represents a priori knowledge of the general system behavior which can be incorporated into the decision. One such estimator is the distance function applied to highway systems. Although in most roads, the cost may not in fact be proportional to distance, it might be a very good estimate.

In general, there are two basic factors to be estimated. The first is the nature of the remaining portion of the route to the destination which occurs as a result of the position in the system following the impending decision. The second is the applicability of some type of system knowledge as an aid in estimating the associated cost of the remainder of the route with the predicted nature of the route. As an example, assume you are on a highway system with two or more alternative routes to the destination. For each alternative, you know the cost of going to the next decision point whether it be the time of travel, gas consumption, etc. You do not, however,have the same information about that part of the route from the next decision points to the destination. You may assume that the remaining route in each case lies on a straight line through the destination. This is an estimate of the first type above. Secondly, you may assume that time of travel, gas consumption, etc. can be obtained from the resulting straight line route. This is an estimate of the second type. If the estimate of the first type is correct, the second could be incorrect due to heavy traffic causing the time estimate to be incorrect or due to excessively mountainous terrain causing the gas consumption to be incorrect, etc. The estimate of the first type could be incorrect due to bodies of water which must be driven around, detours, terrain prohibiting straight roads, etc.

The class of systems presented is modeled as a set of vertices, V, and a set of edges, E[3]. Associated with G = (V,E) is a mapping ϕ such

that for any $e \epsilon E$, $\phi(e) \geq 0$ and represents the cost of following the decision which results in traversing the edge, e.

The purpose of this work is the testing of the estimators and their subsequent evaluation. It is not the purpose of this paper to test the value of the cost proportional to distance assumption. It is convenient in testing to model each system with cost actually being proportional to distance. Such a model allows the testing of the distance estimator for its effectiveness in route determination. In such a case, the basis of the system knowledge, i.e., cost proportional to distance, is valid. Thus the testing of the validity of the estimator is actually testing the applicability of system knowledge of $\phi(e)$ to the graph G = (V,E).

The decision problem involves a given origin $j \epsilon V$ and a goal or destination $k \epsilon V$ and a set of feasible intermediate vertices $I \subset V$ such that for all $i \epsilon I$ there exists an $e \epsilon E$ such that e is incident to i and j. All graph models are assumed to be connected which implies a path exists from every $i \epsilon I$ to k. It is not known, however, what the value of $|P_i| = \sum_{r \epsilon P_i} \phi(r)$ is for any vertex i where $P_i \subset E$ is the set of edges making up the minimum path from i to k. The estimator h(i) is an estimate of the value of $|P_i|$.

In highway systems, one modeling possibility is to establish a mapping α, α: $V \rightarrow R^n$, where R^n is the n-dimensional Euclidean space, i.e., for all $i \epsilon V$, α: $i \rightarrow \underline{x}_i \epsilon R^n$. Thus with feasible intermediate vertex i and destination k with i, $k \epsilon V$, an estimate h(i) is $d(\underline{x}_i, \underline{x}_k)$ where $d(\underline{x}_i, \underline{x}_k)$ is the Euclidean distance between points $\underline{x}_i$ and $\underline{x}_k$. In this paper we will take α such that α: $V \rightarrow R^2$ and, for all adjacent i, $j \epsilon V$ with incident edge $e_{ij} \epsilon E$, $\phi(e_{ij}) = d(\underline{x}_i, \underline{x}_j)$ with $\underline{x}_i$, $\underline{x}_j \epsilon R^2$.

ESTIMATORS

The distance $d(\underline{x}_i, \underline{x}_k)$ along with the estimator $h(i) = (|x_{i_1} - x_{k_1}| + |x_{i_2} - x_{k_2}|)/2$ were presented by Hart, Nilsson, and Raphael[1] as estimates satisfying a required condition for their algorithm A^* in a two-dimensional space. Two additional estimates not satisfying the condition imposed in reference 1 but which are possibilities for other algorithms are

$$(|\underline{x}_{i_1} - \underline{x}_{k_1}| \cdot |\underline{x}_{i_2} - \underline{x}_{k_2}|)/2$$

and

$$[(|\underline{x}_{i_1} - \underline{x}_{k_1}| \cdot |\underline{x}_{i_2} - \underline{x}_{k_2}|)^{1/2}]/2$$

EVALUATION

Each of the four proposed estimators was evaluated for six representative systems. The function h(i) was combined with the known possible costs of the first decision, g(i), from the origin j to ith feasible intermediate vertex.

The six systems used for the evaluation have been illustrated in previous related work and will not be repeated here[4]. The reader is referred to the reference for the descriptions. The four estimators used will be identified as follows:

$$h_1(i) = [(x_{i_1} - x_{k_1})^2 + (x_{i_2} - x_{k_2})^2]^{1/2}$$
$$h_2(i) = [|x_{i_1} - x_{k_1}| + |x_{i_2} - x_{k_2}|]/2$$
$$h_3(i) = [|x_{i_1} - x_{k_1}| \cdot |x_{i_2} - x_{k_2}|]/2$$
$$h_4(i) = [|x_{i_1} - x_{k_1}| \cdot |x_{i_2} - x_{k_2}|]^{1/2}/2$$

Cost was proportional to distance in each of the models and the connectivity of the systems was known, therefore $d(\underline{x}_j,\underline{x}_i)$ is the cost of going from vertex j to vertex i or $g(i) = d(\underline{x}_j,\underline{x}_i)$.

In order to obtain the "optimum" linear combination of $d(\underline{x}_j,\underline{x}_i)$ and h(i), the constants ω_1, ω_2 were used giving:

$$l_i\ (\underline{x}_j,\underline{x}_k) = \omega_1\ d(\underline{x}_j,\underline{x}_i) + \omega_2\ h(i)$$

A successor matrix, $S = (s_{ij})$, is an NxN matrix where if j represents the origin of a route or sequence and k represents the destination of the route or sequence, s_{jk} is the next vertex in the route or sequence following the origin j.

Each estimator was used to generate an apparent successor matrix, T, for each of the six systems for values of ω_1 and ω_2 ranging from 0.0 to 1.0 in steps of 0.1.

The actual testing was performed by using the known successor matrix $\underline{S}$ for each system to compare against the approximate successor matrix $\underline{T}$. Each element of S and T were compared with the number of differences being noted by $\xi(\omega_1,\omega_2)$. The results of this test, excluding trivial cases, showed constant values of $\xi(\omega_1,\omega_2)$ for constant $\frac{\omega_2}{\omega_1}$ which is to be expected. The results were therefore tabulated as $\xi(c)$ with $(\omega_1^2 + \omega_2^2)^{1/2} = 1$ and $c = \tan^{-1}\frac{\omega_2}{\omega_1}$. Thus for the indicated values of ω_1 and ω_2, c is such that, $0 \leq c \leq \pi/2$. The boundaries $\omega_1 = 0$ and $\omega_2 = 0$ indicate rather special conditions which are of no consequence in this investigation and thus c is restricted by $0 < c < \pi/2$.

The "optimum" constants of combination correspond to that value of c given by

$$\underset{0<c<\pi/2}{\text{Min}}\ [\xi(c)]$$

From each plot of $\xi(c)$, a second order least-squares fit was performed and the local minimum of $\xi(c)$, $0 < c < \pi/2$, was obtained from the second order equation.

The results of the evaluation are summarized in Table 1. For each estimator, $h_j(i)$, the minimum number of errors for each system were totaled to give the average percent minimum errors. The values of c at which the minimum value of $\xi(c)$ was obtained were averaged to obtain the "optimum"

value of c. The number of errors with this "optimum" value of c were obtained from the curves of $\xi(c)$. These errors for each system were then totaled to give the average percent errors with "optimum" c.

Estimator	Average Percent Minimum Errors	Average Percent Errors with "optimum" c	"optimum" c
$h_1(i)$	12.6	13.8	0.62
$h_2(i)$	15.4	18.2	0.74
$h_3(i)$	34.1	35.2	0.74
$h_4(i)$	36.9	37.3	0.83

Table 1. Error Percentages

SUMMARY

The use of estimators for the route from a feasible intermediate vertex to the destination is important to any heuristic method of solution. The estimates of the first type are considered in this investigation. Four estimators are used in conjunction with a known quantity involved from the origin to the feasible intermediate vertex to make up a policy equation. It is found that for the systems studied it is possible to obtain a pair of constants which form the linear combination which is very nearly optimum for the six given systems.

It is found for the estimates of the first type that the distance estimator gives the best results. It is the authors' belief that the method of comparison used in this paper can be used as a basis for further work in this area. The phenomena being tested is complicated to say the least. Only through additional work can the necessary conclusions be drawn with certainty. It is felt that the results of this work are sufficiently correlated to encourage future work.

ACKNOWLEDGEMENT

The authors wish to acknowledge the Computer Center of the University of Pittsburgh for making their facilities available and for the support provided under the National Institute of Health Grant FR-00250.

REFERENCES

1. P. E. Hart, N. J. Nilsson, B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths", IEEE Transactions on Systems Science and Cybernetics, Vol. 4, No. 2, pp. 100-107.

2. Mickle, M. H., Sze, T. W., Rathbone, D. E., "Neighborhood Storage", IEEE Transactions on Systems Science and Cybernetics, Vol. 4, No. 2, p. 138-144.

3. Seshu and Reed, Linear Graphs and Electrical Networks, Addison-Wesley, 1961.

4. Mickle, M. H., "On a Policy Equation and Contours with Neighborhood Storage", Proceedings of the First Houston Conference on Circuits, Systems, and Computers, Houston, May 1969.

A SYSTEM THEORY FRAMEWORK FOR MODELLING SECTORAL PRICES IN AN INTERREGIONAL ECONOMY

PREM S. SATSANGI
Electrical Engineering Department, Indian Institute of Technology
Hauz Khas, New-Delhi 29, India

JACK B. ELLIS
Environmental Studies and Electrical Engineering, University of
Waterloo, Waterloo, Ontario, Canada.

ABSTRACT

This paper extends a physical system theory philosophy, which is a generalization of network theoretic concepts, to the modelling of prices in a national economy. A "pilot" study for the Canadian economy at a level of disaggregation corresponding to 10 industrial sectors in each of 5 geographical regions is briefly reported.

INTRODUCTION

There has long been a need for developing a general theoretical framework for economic systems research which imbeds dynamical, structural, and spatial aspects in system models, and emphasizes descriptive behavioral modelling with predictive powers. The national economy is identified here as a collection of interconnected components and the approach taken is that of a system theory construct. Each economic component is modelled in terms of two fundamental complementary variables, a through variable Y - the flow of "commodities" valued in constant dollars, and an across variable X - the unit price index. The implications of the choice of a general "hybrid" form of component models in terms of the computer-oriented procedures of the systems modelling theory are discussed.

In the specific context of the objectives of this paper, the pertinent behavioral characteristics of each N-terminal economic component are completely specified by a set of N-1 terminal equations of the form of "generalized drivers" and a terminal graph. Typically, the terminal characteristics embody economically meaningful parameters; inter-industry technical input-coefficients and capital-coefficients, interregional trade coefficients, depreciation rates, monetary interest rates, capital-output and labour-output ratios, "vintage" capital and labour embodied technological progress factors.

Exogenous variables are industrial final-demands, interregional transportation prices, employed labour-force and wages or labour-income per unit output. The system graph is obtained operationally by uniting the vertices of the terminal graphs used to model the characteristics of the components in a one-to-one correspondence with the interconnection of the components in the economic system. A reduced form of the graph is shown in Figure 1. The system graph furnishes interconnection constraint equations which together with the component terminal equations characterize the system completely, such that it can be described by means of a discrete-time state-space model. This development is consistent in the sense that there is a balance of payments in the closed economic system.

The process of system-modelling for large-scale systems has been rationalized further by suitably identifying and modelling subassemblies of "interregional" and "inter-sectoral" types and then recombining these models (i.e., by tearing and reconstruction). It is believed that the

framework is general enough so as to embrace practically all present "input-output" models in the economic literature. Specifically, specialization to Isard-Leontief, Moses, and Leontief-Strout interregional models (1,2,3,4) has been demonstrated and several desirable refinements made. Price and flow models have been empirically implemented for the purposes of a modest "pilot" study of the Canadian economic system over the period 1949-1976 at a level of 10 industrial sectors in each of the 5 regions of Canada. As a by-product of the empirical implementation, "new" investment calculations are made possible and there results a capability to estimate technical bias for each of the Canadian industries considered.

This paper presents a summary of the development and results of the price model for the Canadian economy.

PRICE MODEL FORMULATION

The pertinent behavioral characteristics of each N-terminal economic component are completely specified by a set of N-1 "terminal equations" and a "terminal graph" with N-1 edges. Each "terminal" represents a functional interdependency in economic terms, or an actual connection point as in transport terms. The relevant economic components are classified as follows:

(i) Transportation edges characterized as ordinary 2-terminal across drivers with interregional transport price as a specified function of time.

(ii) Regional intermarket edges characterized as ordinary 2-terminal across-drivers with intermarket transport price in some reference sector as a specified function of time.

(iii) Dynamic, regional market and production components, representing local intersectoral transactions and prices, characterized as multi-port modified couplers.

(iv) Sectoral, interregional and intermarket trade, and demand components, representing interregional receipts, sectoral demands, and prices, characterized as multi-terminal augmented couplers.

(v) Exogenous demand edges characterized as ordinary 2-terminal through drivers.

The system graph G with v vertices and e edges is obtained by uniting the vertices of the terminal graphs in a 1:1 correspondence with the interconnection of the components in the economic system. This furnishes (e-v+1) linearly independent circuit equations which, in general, express price-cost equality; between exogenous demand or intermediate demand and total regional demand; between intra-regional supply and production; between interregional supply on one hand and production plus interregional transport on the other; and between to-and-fro intermarket shipments and to-and-fro interregional transportation. These interconnection price-constraint equations together with the partial set of component terminal equations in prices describe a general price state-model of the national economic system which permits complete solution for all price-indices in the system. This model is characterized dynamically by the following difference equation form:

$$X(t) = \left\{(\theta^1 + \phi^1)\ T^1 - I\right\}^{-1} \left\{\phi^1\ T^1\ X(t-1)\ -Xe(t)\right\} \quad .....(1)$$

where,

$$Xe(t) = Xw(t)\ + \theta^1\ Xp(t)\ + \phi^1 \nabla Xp(t)$$

such that the vectors and matrices (each of the size mn; m being the number of regions and n the number of sectors in the economy) have the following interpretation:

θ and $\emptyset$ are the operational price-index parameter-matrices which, in general, are time-variable and typically embody economically meaningful parameters; interindustry technical input coefficients and capital coefficients, depreciation rates, lagged monetary interest rates, etc.

T is the "scattered" trade coefficient matrix
I is the identity matrix
X is the production price-vector
Xw is the vector of exogenous price-index generator
(Typically, Xw's represent labour-income per unit output plus "extra" entrepreneurial profits per unit output)
Xp is the aggregate transport price-vector determined by interregional trade coefficients and exogenous interregional transport prices.
∇ is the backwards difference operator and Primes over matrices denote their transpose.

METHOD OF SOLUTION

It is well known that the dynamic Leontief model (a special case of (1)) is plagued by the highly ill-conditioned nature of the "B-matrix", some rows of which are almost zero. This is adversely reflected in the direct implementation of model (1) on the digital computer. However, this difficulty is bypassed in the following operational version of the model which employs Neumann's series expansion up to 2 (or 3) terms only:

$$X(t) = (I-\theta^1(t)\,T^1)^{-1}\,Xe(t) + (I-\theta^1(t)\,T^1)^{-1}(\emptyset^1(t)\,T^1)\,\nabla\{(I-\theta^1(t)\,T^1)^{-1}\,Xe(t)\} + \ldots\ldots\ldots\ldots \quad (2)$$

EMPIRICAL IMPLEMENTATION

It has been shown (5) at the operational level that a sufficient degree of regionalization of Canadian national accounts can be arrived at to establish, without too much expense and for each programming region, a good enough approximation of regional accounts consistent with the national accounting from which they are drawn.

A general computer programme "PRIDEX" has been written which generates price-index solutions for the large scale Canadian economic system in the period 1949-1976. Verification and accuracy analyses for the simulation period (1949-1965) show the good qualities of the model whose conditional forecasts for the future are also plausible.

While limitations on space prevent us from presenting the detailed results, these can be obtained from the authors. Figures 2, 3 and 4 presents a summary of price index computations versus actual indices for three selected sectors. Figure 1 indicates a form of the system graph which has been reduced for graphical clarity.

REFERENCES

1. W. Isard, et al., Methods of Regional Analysis: An Introduction to Regional Science, John Wiley, New York, 1960.

2. W. Isard and J.H. Cumberland, "A Synthesis of Operational Methods in Regional Analysis", in Regional Economic Planning, (Editor: W. Isard and J.H. Cumberland), Organization for Economic Cooperation and Development, Paris, 1961.

3. W.W. Leontief and A. Strout, "Multiregional Input-Output Analysis", in Structural Interdependence and Economic Development (Editor: T. Barna), St. Martin's Press, New York, 1963.

4. L.N. Moses, "The Stability of Interregional Trading Patterns and Input-Output Analysis", American Economic Review, Vol. 45, No. 5, Dec. 1955, pp. 803-832.

5. P.S. Satsangi, "A Physical System Theory Modelling Framework Generalized For Large Scale Economic Systems", unpublished Ph.D. thesis, Department of Electrical Engineering, University of Waterloo, 1969.

FIGURE 1: REDUCED SYSTEM GRAPH

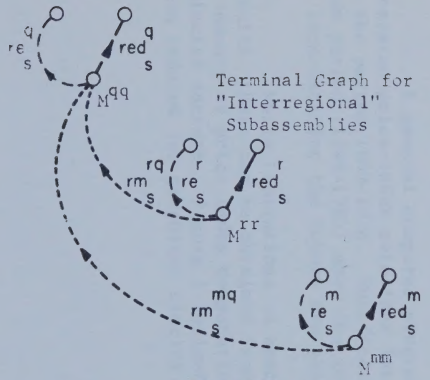


Figure 2-7

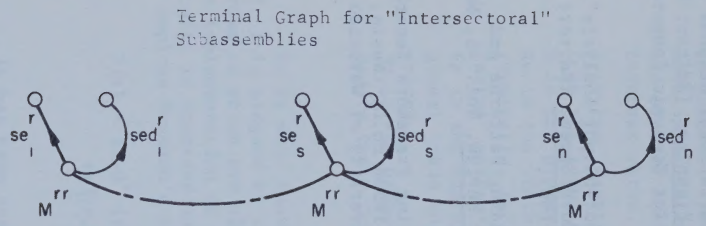


Figure 2-8

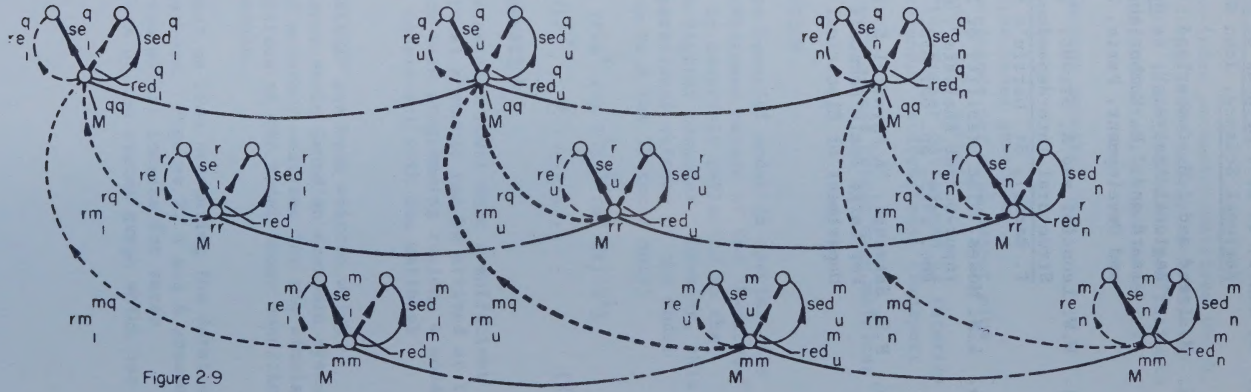
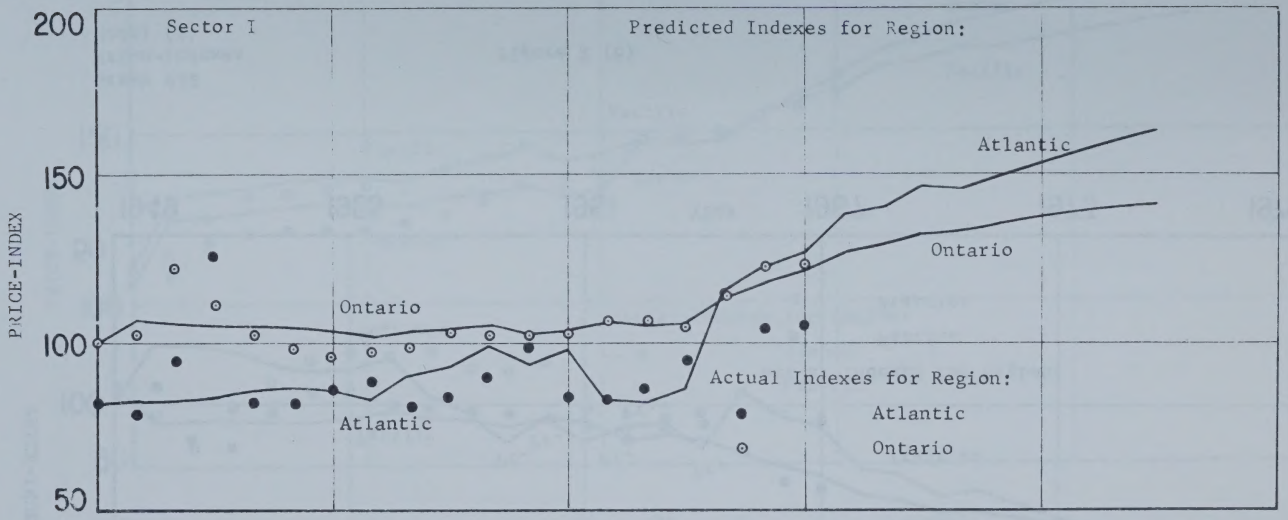


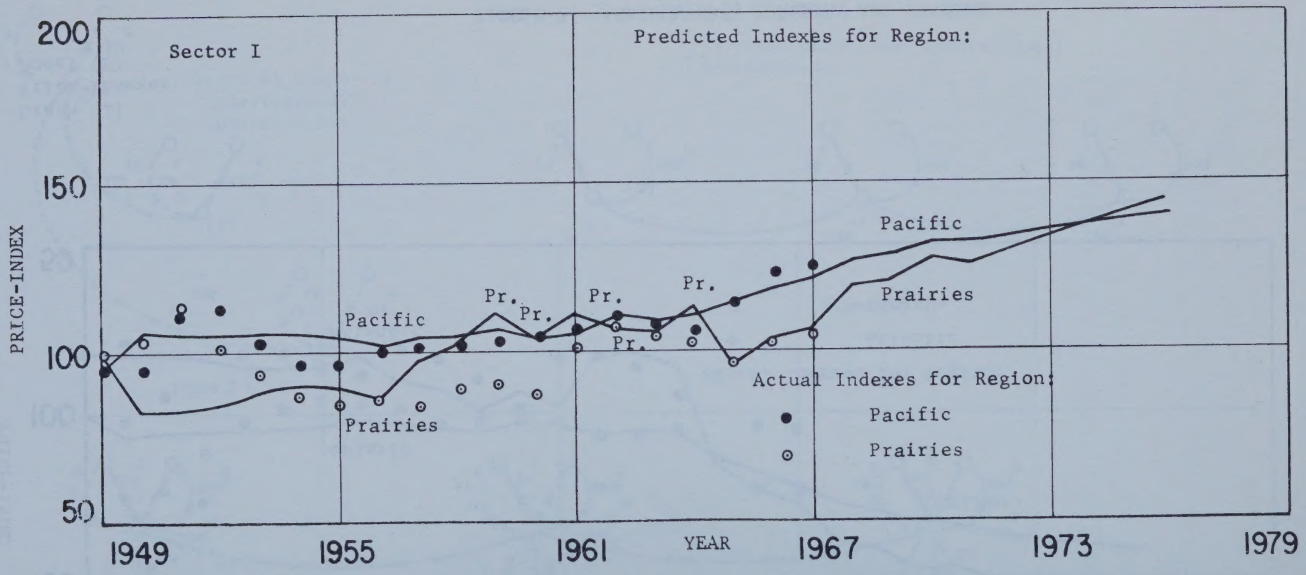
Figure 2-9



(a)

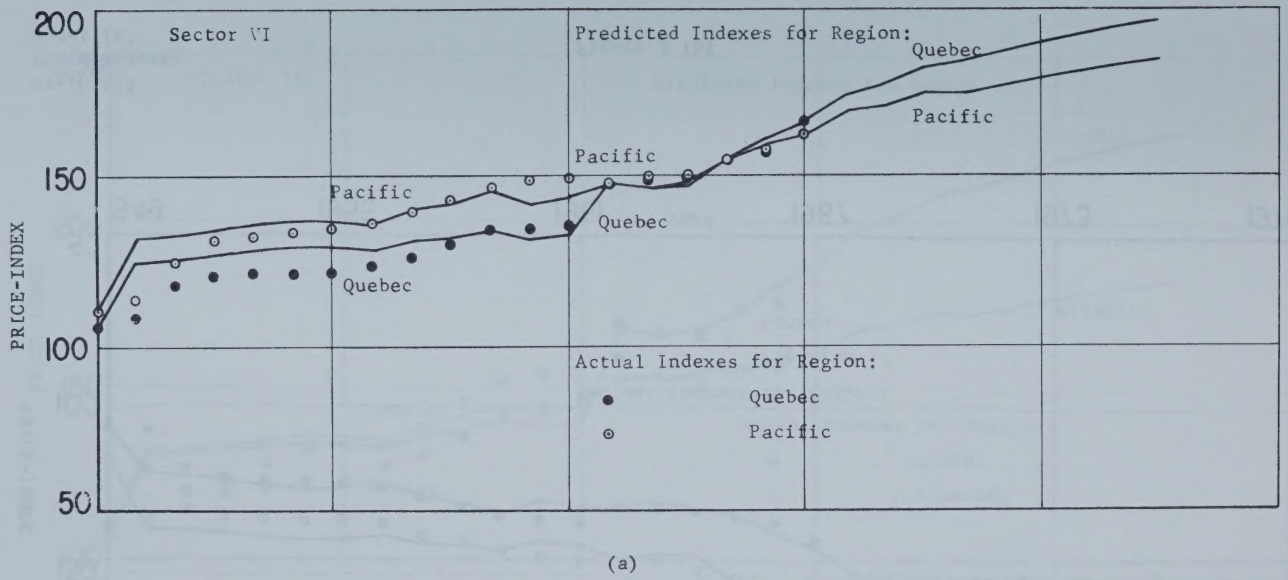
Graph #21
Price-Indexes
Model (B)

FIGURE 2: AGRICULTURE; FORESTRY AND FISHING



Graph #22
Price-Indexes
Model (B)

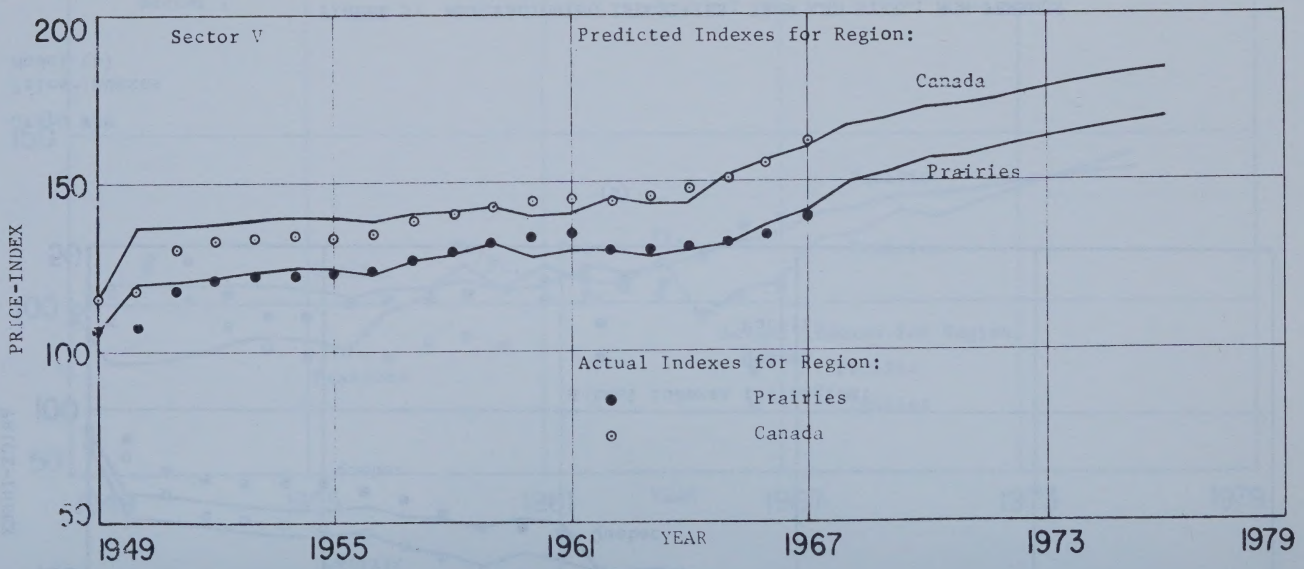
Figure 2 (b)



Graph #36

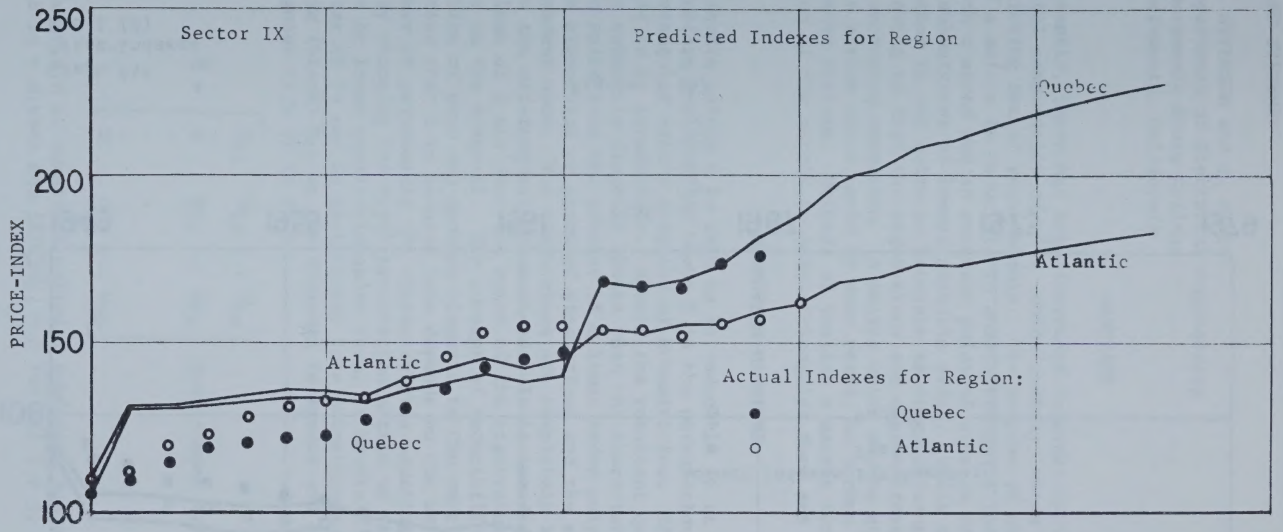
Price-Indexes
Model (B)

FIGURE 3: MANUFACTURING INDUSTRIES; IRON AND STEEL; NON-FERROUS METAL; TRANSPORTATION EQUIPMENT; ELECTRICAL APPARATUS; OTHERS



Graph #35
Price-Indexes
Model (B)

Figure 3 (b)



(a)

Graph #45
Price-Indexes
Model (a)

FIGURE 4: TRANSPORTATION, STORAGE AND TRADE

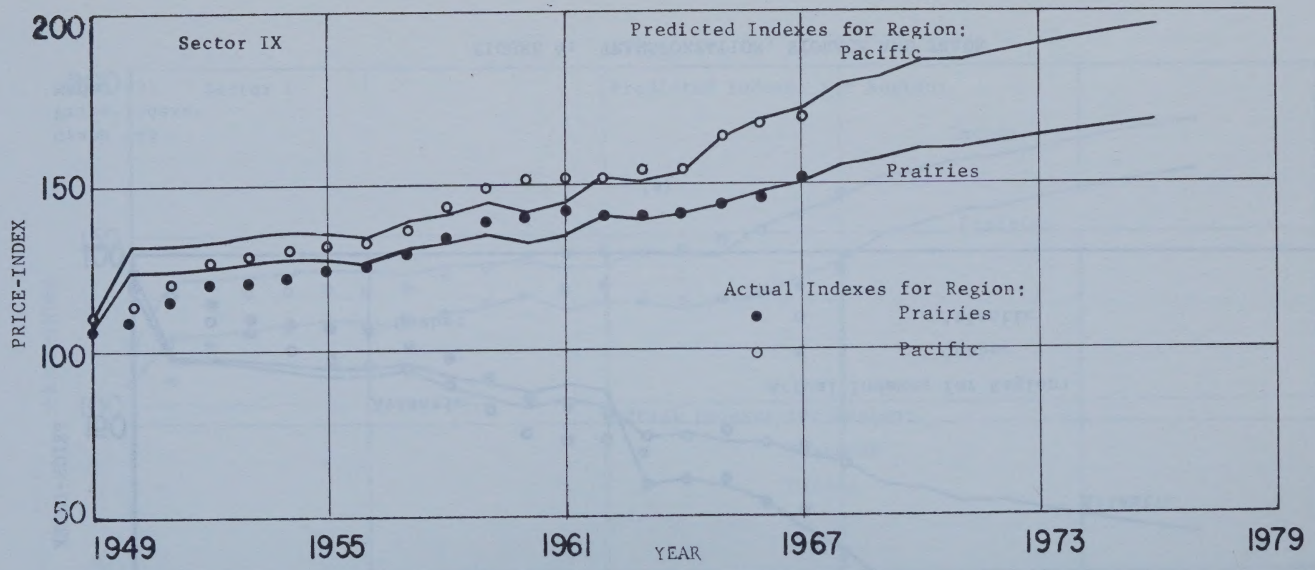


Figure 4 (b)

Graph #46
Price-Indexes
Model (B)

REDUCIBLE MATRICES AND THEIR APPLICATION TO COMPUTER ANALYSIS OF LINEAR TIME INVARIANT SYSTEMS*

D. SYVERSON and J. SIMES
Department of Electrical Engineering
Sacramento State College
Sacramento, California

ABSTRACT

Recently, there has been increased interest in fast and accurate methods of digital computation of e^{AT}. Unfortunately, these two goals often present a conflicting set of requirements. The purpose of this paper is to suggest that if a matrix is reducible to quasi-triangular form, then it is desirable from both a speed and an accuracy point of view to reduce the matrix into a set of submatrices of lowest possible order. While the eventual solution of this method is not quite as accurate as diagonalizing the matrix, it does avoid having to find the eigenvalues and eigenvectors of the matrix, normally a time consuming operation. A matrix which may be quasi-triangularized describes a system which could by some means, be broken up into a cascade of lower order systems. Methods of testing a matrix for reducibility and the reduction of the matrix to quasi-triangular form are discussed.

REDUCIBLE MATRICES

A square matrix A is said to be reducible if it satisfies the similarity transform $M=P^{T}AP=P^{-1}AP$ where P is the permutation matrix and M is the quasi-triangular matrix composed of submatricies. If no such matrix exists, A is said to be irreducible. Should the resultant quasi-triangular matrix M have m submatrix diagonal blocks then the characteristic polynomial of A has been split into the product of m lower order polynomials thereby isolating each eigenvalue or group of eigenvalues and the system parameters they are dependent upon. The eigenvalues are completely independent of the elements in the off-diagonal submatrices. It is immediately apparent that the eigenvalues of A are just the union of the eigenvalues of the resulting submatrices on the diagonal. The concept of reducibility only depends upon the disposition of zero and nonzero elements in the matrix A. The number of submatrices that A is reduced into depends on the matrix A, in particular the number of polynomials the characteristic equation of A can be factored into. By choosing the right permutation matrix we can reduce A into either an upper or lower quasi-triangular form. A matrix M is called quasi-upper triangular if it can be partitioned into submatrices M_{ij} such that each submatrix (block) M_{ii} on the diagonal is square and each of the M_{ij} below the diagonal ($i > j$) is zero.

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ 0 & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & M_{nn} \end{bmatrix} \quad \text{quasi-upper triangular}$$

Such a matrix M is sometimes called a subdirect sum of the diagonal blocks M_{ii}. It is a direct sum iff. $M_{ij} = 0$ for all $i \neq j$.

*This research was supported by NSF Grant #GK-5200.

If the unit vector e_j is the vector with a 1 in the jth position and zeros elsewhere, then a permutation matrix is a nxn matrix of the form $P = [e_{i1}e_{i2} \ldots e_{in}]$ where i1 i2 ... in is a permutation of 123 ... n. The effect of postmultiplying (premultiplying) a matrix A by P is to interchange the columns (rows) of A according to the permutation. A permutation matrix is necessarily orthogonal since it is composed of n linearly independent unit vectors. The similarity transform of A of the form P^TAP simply has the effect of rearranging the elements of A; rows and columns being permuted correspondingly.

If a graph is made from the matrix A such that for each nonzero element a_{ij} of A there is a "directed line" connecting node i to node j, then the result is called the "directed graph" or digraph of A. A set of nodes in the graph is said to be strongly connected iff. there is a directed line or sequence of directed lines connecting each node to every other node. The elements a_{ij} representing this strong connection are dependent only upon each other and cannot themselves be separated. These elements form a submatrix block of the reduced A matrix. This submatrix block is irreducible. In other words, a matrix is irreducible iff. every node in the graph is reachable by a directed line or sequence of directed lines, from every other node. A reducible matrix has a directed graph which can be broken up by inspection into strongly connected subgraphs and other elements. These strongly connected subgraphs represent the irreducible submatrix blocks which always lie on the diagonal of the reduced matrix form M. It is interesting to note that the elements in an irreducible submatrix will consist only of the elements whose indices are the permutation of the node numbers in a strongly connected subgraph. It should be mentioned that sometimes reduction is implied by inspection of the directed graph, but the matrix is already in reduced form. This condition can be detected by determining the strongly connected nodes, then taking all elements whose indices are a permutation of these node numbers and see if they already form a submatrix on the diagonal of the given matrix. The irreducible submatrix blocks are determined from the "reachability" matrix. The reachability matrix is easily determined by use of Warshall's algorithm[5] if a computer is available, otherwise the following method is advised:

1. Form the boolean matrix D by replacing all $a_{ij} \neq 0$, and all a_{ii} with one's.
2. Perform the boolean multiplication of D to obtain D^{n-1} where n is the dimension of D, the resultant matrix is called the "reachability" matrix R. The name "reachability" is used because if $R = r_{ij}$, we will have $r_{ij} = 1$ iff. there is a directed path from node i to j. Boolean multiplication of two boolean matrices means the ordinary matrix product in which the boolean rules for product (intersection) and sum (union) are used.

If the reachability matrix contains all one's then the matrix A is irreducible, however, if it contains one or more zeros, then the original matrix is reducible.

The following algorithm (see 2) determines the irreducible submatrix blocks on the diagonal:

1. Determine the reachability matrix R.
2. Let S_1 be the collection of all indicies i such that there is a one in the $i^{\underline{th}}$ place of the first row of R and in the $i^{\underline{th}}$ place of the first column of R. Delete from R all rows and columns whose indicies are in the set S_1, but preserve the original row

and column indicies from R in the resulting submatrix of R.

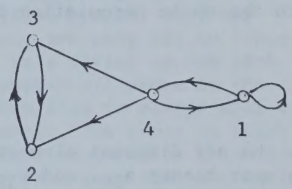
3. Repeat step 2 on the resulting submatrix, etc., obtaining the collection of sets of indicies S_2, S_3, ... S_k.
4. Let M_i be the n_i by n_i submatrix formed by the elements of matrix A from the intersection of the rows and columns listed in the set S_i. The original matrix has now been reduced to a matrix M whose submatrix blocks on the diagonal are irreducible and are M_i. The set of eigenvalues of A is the union of sets of eigenvalues of the submatrices M_i.

From these submatrices on the diagonal of M, the permutation matrix P can be determined such that $M = P^T AP$ thereby giving the location of the elements in the off diagonal submatrices.

Example: As an example in reducing a reducible matrix and determining the permutation matrix necessary to accomplish this, consider the following A matrix:

$$\begin{bmatrix} a_{11} & 0 & 0 & a_{14} \\ 0 & a_{22} & a_{23} & 0 \\ 0 & a_{32} & 0 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

System Matrix A



Directed Graph

The directed graph for this matrix is shown. For this A matrix we have the following boolean matrix from which we obtain the reachability matrix.

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

BOOLEAN MATRIX D

$$R = D^3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

REACHABILITY MATRIX

Since matrix R contains zeros, it is reducible. Matrix R gives sets $S_1 = \{1,4\}$ and $S_2 = \{2,3\}$, from which the submatrices M_1 and M_2 on the diagonal of M are obtained.

$$M_1 = \begin{bmatrix} a_{11} & a_{14} \\ a_{41} & a_{44} \end{bmatrix}, \qquad M_2 = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & 0 \end{bmatrix}$$

The zero in M_2 must correspond to the element a_{33}. The reduced matrix may be found in any one of the following forms depending upon which permutation matrix we use.

$$M = \begin{bmatrix} M_1 & M_3 \\ 0 & M_2 \end{bmatrix}, \begin{bmatrix} M_1 & 0 \\ M_3 & M_2 \end{bmatrix}, \begin{bmatrix} M_2 & M_3 \\ 0 & M_1 \end{bmatrix} \text{ or } \begin{bmatrix} M_2 & 0 \\ M_3 & M_1 \end{bmatrix}$$

The location of the diagonal elements in each submatrix will determine the location of the off diagonal elements in that submatrix. Assume that the reduced form desired is

$$M = \begin{bmatrix} M_1 & M_3 \\ 0 & M_2 \end{bmatrix},$$

and that we should like the location of the diagonal elements of M_1 and M_2 to be the same as given above.

$$M = \left[\begin{bmatrix} a_{11} & a_{14} \\ a_{41} & a_{44} \end{bmatrix} \begin{bmatrix} & \\ & M_3 \end{bmatrix} \right.$$
$$\left. \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & 0 \end{bmatrix} \right]$$

Starting in the upper left hand corner, and observing diagonal elements only, a_{11} stays a_{11}. The next element on the diagonal is a_{44} and it becomes a_{22}. In the next submatrix a_{22} is to become a_{33} and a_{33} is to become a_{44}, therefore the whole permutation is now given as

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \uparrow & \uparrow & \uparrow & \uparrow \\ 1 & 4 & 2 & 3 \end{array}$$

For the off diagonal elements, a_{14} must become a_{12} and a_{41} must become a_{21}, a_{23} must become a_{34}, and a_{32} must become a_{43}, as specified by the permutation. These elements are therefore already arranged in the correct positions and no rearrangement of them will occur in the reduced matrix. The permutation matrix given by this permutation may be derived by permuting either the rows or the columns of an identity matrix. If the rows are permuted then the reduced form will be given by $M = PAP^T$ and if the columns are permuted the reduced matrix form will be given by $M = P^TAP$. We want the form to be given by P^TAP therefore the permutation matrix is given by

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$M = P^TAP = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & 0 & 0 & a_{14} \\ 0 & a_{22} & a_{23} & 0 \\ 0 & a_{32} & 0 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{14} & 0 & 0 \\ 0 & 0 & a_{22} & a_{23} \\ 0 & 0 & a_{32} & 0 \\ a_{41} & a_{44} & a_{42} & a_{43} \end{bmatrix} = \left[\begin{bmatrix} a_{11} & a_{14} \\ a_{41} & a_{44} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ a_{42} & a_{43} \end{bmatrix} \right.$$
$$\left. \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & 0 \end{bmatrix} \right]$$

This gives the characteristic equation $(\lambda^2-a_{22}\lambda-a_{32}a_{23})(\lambda^2-(a_{44}+a_{11})\lambda+a_{11}a_{44}-a_{41}a_{14}) = 0$ which can be solved directly as two independent systems.

PERMUTATION MATRICES

The permutation matrix is determined uniquely by the order of the diagonal elements in the submatrix blocks on the diagonal of the reduced form M. Many permutations are made possible simply by rearranging the diagonal elements of these submatrices or rearranging these submatrices themselves. However, only a certain number of these permutations will reduce matrix A to quasi-triangular form, the rest will merely arrange the elements of A in the positions as specified by the desired submatrices on the diagonal leaving non-zero elements above and below this quasi-diagonal. One possible method for determining one or all of the permutation matrices which give quasi-triangularization, is to determine the maximum connectivity signal flow graph or its matrix representation. By maximum connectivity, it is meant that the new signal flow graph will contain all branches of the original flow graph and all other possible branches such that this graph contains no loops (cycles) except all of those in the original graph. Once the maximum connectivity signal flow graph is obtained, one only has to trace out all possible paths from the source node (may be a cycle) to the sink node (may be a cycle) such that the path taken touches all nodes in the graph only once (called a proper path) and that the flow is always with (or against) the direction of the directed line segments. The order in which the nodes are traversed, in one of these paths, is a permutation which gives quasi-triangularization. If the flow through the path goes with and against the directed line segments, then the resulting permutation is one of those which fails to give quasi-triangularization. Flow may also go from sink to source. For low order systems the above process may be accomplished by inspection of the original digraph. In higher order systems the maximum connectivity graph, and even the paths through it, cannot be easily determined by inspection. Therefore, it is desirable to have an algorithm to accomplish this.

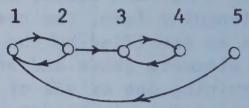
One approach to this problem is to treat the system as a finite state machine, then the maximum connectivity matrix and all correct proper paths through it can both be determined by obtaining the (n-1)-order transition (maximum connectivity) matrix for the given nth order system. This matrix is nothing more than the (n-1)th power of the connectivity matrix obtained by setting all elements representing proper loops (cycles) in the matrix to zero prior to each multiplication by matrix A which also has the elements in it representing proper cycles set to zero. A path is called a proper cycle if the path traversed, as represented by the matrix element or elements, has indices $i, m_1, m_2, \ldots, m_{k-1}, j$, which are not all distinct. The resultant matrix $A^{(n-1)}$ obtained this way contains only nonzero elements which are the correct paths through the maximum graph. Setting all proper cycles to zero before each matrix product, will result in all diagonal elements being set to zero, as these elements represent all the cycles not existing in the original graph. Note that A^2 will contain all possible cycles of length two on its diagonal, A^3 contains all possible cycles of length 3 on its diagonal, etc. The proper cycles set equal to zero give a set of necessary conditions which must be satisfied in order to obtain the maximum graph (matrix). Delete all conditions of this set which contain only elements of the original flow graph. Elements which are forced to be zero are used in the next product to simplify calculations. These conditions will usually show that this graph (matrix) is not unique. It should be mentioned that multiplication of higher-order transition matrices is the

same as the multiplication of ordinary matrices except that the order of the factors in each product $a_{iu}a_{uj}$ must be preserved: $a_{iu}a_{uj}$ is not equivalent to $a_{uj}a_{iu}$. The power may be represented by the recursive formula of $A^{K+1} = A^K A$, $K=1,2,\ldots,n-2$. In the example that follows, the matrices are shown after all proper cycles are set to zero. The conditions which are necessary for the new proper cycles to equal zero are listed as they are found. All matrix elements are included to clarify the process.

Example:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

System Matrix A



Digraph

$$A^2 = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & 0 & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & 0 & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & 0 & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & 0 \end{bmatrix}^2$$

$a_{15}, a_{32}, a_{31} = 0$ by inspection

$$= \begin{bmatrix} (0) & (0) & (a_{12}a_{23}+a_{14}a_{43}) & (a_{12}a_{24}+a_{13}a_{34}) & (a_{14}a_{45}) \\ (0) & (0) & (a_{21}a_{13}+a_{24}a_{43}) & (a_{21}a_{14}+a_{23}a_{34}) & (a_{24}a_{45}) \\ (0) & (0) & (0) & (0) & (a_{34}a_{45} \\ (a_{45}a_{51}) & (a_{45}a_{52}) & (a_{45}a_{53}) & (0) & (0) \\ (a_{52}a_{21}) & (a_{51}a_{12}) & (a_{51}a_{13}+a_{52}a_{23}+a_{54}a_{43}) & (a_{51}a_{14}+a_{52}a_{24}+a_{53}a_{34}) & (0) \end{bmatrix}$$

Proper cycles give: $a_{13}a_{31} = a_{14}a_{41} = a_{15}a_{51} = a_{23}a_{32} = 0$

$a_{24}a_{42} = a_{25}a_{52} = a_{35}a_{53} = a_{45}a_{54} = 0$

Since: $a_{51} \neq 0, \quad a_{23} \neq 0, \quad a_{13} \neq 0$

Then conditions are: $a_{31} = a_{15} = a_{32} = 0$

$a_{14}a_{41} = a_{24}a_{42} = a_{25}a_{52} = a_{35}a_{53} = a_{45}a_{54} = 0$

Finally,

$$A^3A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & (a_{52}a_{21}a_{14}a_{43}+a_{51}a_{12}a_{24}a_{43}) & (a_{52}a_{21}a_{13}a_{34}+a_{51}a_{12}a_{23}a_{34}) \end{bmatrix} = A^4$$

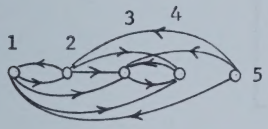
The total set of conditions to give this final form of A^4 is:

$$a_{15} = a_{42} = a_{25} = a_{31} = a_{45} = a_{41} = a_{32} = a_{35} = 0$$

$$, a_{24},\ a_{14}a_{52},\ a_{52},\ a_{24}a_{52},\ a_{53},\ a_{14} \text{ may or may not} = 0$$

$$a_{54} \text{ not specified } \therefore \text{ may or may not} = 0$$

To obtain maximum graph let arbitrary elements = 1



Resulting Maximum Graph

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 \\ 0 & 0 & a_{43} & a_{44} & 0 \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

Maximum Connectivity Matrix

From matrix A^4 we see that there are four possible forward paths. These paths in reverse order give four more paths (reverse paths). The resulting permutations for these eight paths are: 52143, 34125, 51243, 34215, 52134, 43125, 51234 and 43215. Further investigation will show that all permutation matrices derived from these permutations will yield the desired quasi-triangular form.

APPLICATION

Consider the linear time invariant system described by

$$\dot{\underline{x}}(t) = A\underline{x(t)}$$

$$y(t) = C\underline{x(t)}$$

As is well known,

$$x(T) = e^{AT}x(o)$$

and a digital computer solution is often found by using the recursion formula

$$x[(n+1)T] = e^{AT}x(nT)$$

For this reason, methods of calculation of e^{AT} are of importance. (See 3, 4, 6.)

It is desirable, from both a speed and an accuracy point of view, that A be in diagonal form. However, this requires complete knowledge of the eigenvalues, normally a time consuming, and not particularly accurate operation. Therefore, one seeks other forms of the A matrix which yield as rapid and as accurate solutions as are possible without solving for the eigenvalues. One such form is the quasi-triangular form. Consider Fig. 1.

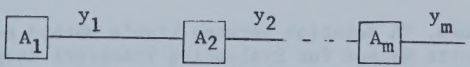


Fig. 1 Cascaded System

If

$$\dot{\underline{x}}_i = A_i\underline{x}_i, \quad \underline{y}_i = C_i\underline{x}_i$$

then the complete solution is given by

$$\dot{\underline{x}} = A\underline{x}$$

$$\underline{y} = [C_m, \ 0 \ 0 \ \ldots \ 0]\underline{x}$$

where

$$A = \begin{bmatrix} [A_m][C_{m-1}] & & \\ & \ddots & \\ & & [A_2][C_1] \\ \bigcirc & & [A_1] \end{bmatrix}$$

This A matrix is quasi-upper triangular. Likewise, any matrix which is quasi-upper triangular, may represent a cascaded system. In the general case, some of the individual boxes might have feed-forward paths around them, and all of the states of any individual system might feed forward. Therefore, it is possible to completely fill the upper half of the matrix.

The quasi-triangular form has several characteristics desirable for computer operation. First, if A is quasi-triangular, then A^n is also quasi-triangular, and of the same form. Therefore, the number of operations necessary to raise the A matrix to a power can be greatly reduced. Second, if all of the system time constants are less than one, then the numbers associated with a quasi-triangular matrix are smaller in magnitude than those of the companion form. Therefore, a Taylor series expansion of A to calculate e^{AT} converges more rapidly. Third, the eigenvalue sensitivity of e^{AT} in quasi-triangular matrices is in general smaller than that of the companion form, although it is not as small as the diagonal form.

CONCLUSION

Methods of testing a matrix for reducibility are discussed. The application of the permutation matrix, reachability matrix, and maximum connectivity matrix are explained. It is shown that reducible matrices may be represented by cascaded linear time invariant systems. Some properties of quasi-triangular matrices are discussed.

REFERENCES

1. N. D. Francis, "Sensitivity Analysis in Multivariable-System Design," Electronic Letters, Vol. 3, No. 9, Sept. 1967, pp. 404-405.

2. F. Harary, "A Graph Theoretic Method for the Complete Reduction of a Matrix with a View Toward Finding Its Eigenvalues," Journal of Mathematics and Physics, Vol. 38, 1959, pp. 104-111.

3. M. Lion, "A Novel Method of Evaluating Transient Response," Proc. IEEE, Vol. 54, Jan. 1966, pp. 20-23.

4. E. J. Mastascusa, "A Relation Between Liou's Method and the Fourth Order Runge-Kutta Method for Evaluating Transient Response," Proc. IEEE, Vol. 57, May 1969, pp. 803-804.

5. S. Warshall, "A theorem on Boolean Matrices," Association for Computing Machinery Journal, Vol. 9, No. 1, Jan. 1962, pp. 11-12.

6. E. J. Mastascusa, "A Method of Calculating e^{AT} Based on the Caley-Hamilton Theorem," to be published as a proceedings letter, Proc. IEEE.

A NEW APPROACH TO THE EVALUATION OF THE TRANSITION MATRIX

ILIA KAUFMAN
Department of Systems Design
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

This paper describes a new algorithm for the evaluation of the transition matrix (or any other analytical function) of a given square matrix. The method is based on a recently developed closed form expression for functions of companion matrices. It is also indicated that the algorithm suggested in this paper involves substantially less computational effort than the commonly used diagonalization procedure.

It is well known that any analytical function g of a square matrix A can be represented as

$$g(A) = R\, g(B)\, R^{-1} \tag{1}$$

where A and B are similar matrices, that is,

$$A = R\, B\, R^{-1} \tag{2}$$

Obviously, if B is a diagonal matrix,* then g(B) can be easily evaluated, yielding through (1) the desired function g(A). Observe that this approach (known as the diagonalization procedure) involves the computation of the complete eigensystem of the matrix A. In addition, since a general real matrix A may have complex eigenvalues, the whole similarity transformation (2) should be expected to be complex. Thus, the evaluation of the functions of real matrices through the diagonalization procedure involves "unnecessary" complex matrix operations.

These unnecessary complex matrix operations can be eliminated if instead of diagonalizing the matrix A, one reduces it to its companion form. Such a "trade" was made possible through the development of the following efficient and accurate algorithm for the evaluation of functions of companion matrices.

Let B be the following companion matrix

$$B = \begin{bmatrix} 0 & 1 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 1 \\ -a_n & -a_{n-1} & \cdot & \cdot & \cdot & -a_1 \end{bmatrix} \tag{3}$$

whose characteristic polynomial is

$$f(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \tag{4}$$

and whose eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_n$ are all distinct.

---

* Matrices with distinct eigenvalues are necessarily similar to diagonal matrices. Since the presence of multiple eigenvalues implies an infinite precision of measurements as well as an infinite precision of computations, it can be assumed without great loss of generality that matrices describing real physical systems are similar to diagonal matrices with distinct eigenvalues.

It has been shown[1,2] that if P represents some analytical function g of B, that is P = g(B), then the elements of P can be computed as follows

Step 1: Compute

$$p_{i,n} = \sum_{l=1}^{n} \frac{g(\lambda_l)}{f'(\lambda_l)} \lambda_l^{i-1} \qquad i=1,2,\ldots,n \tag{5}$$

where

$$f'(\lambda_l) = \prod_{j \neq l} (\lambda_l - \lambda_j) \tag{6}$$

Step 2: Compute

$$p_{i,j} = p_{i+1,j+1} + p_{i,n} a_{n-j} \qquad \begin{matrix} j=n-1,n-2,\ldots,1 \\ i=1,2,\ldots,j \end{matrix} \tag{7}$$

Step 3: Compute

$$p_{i,1} = -a_n p_{i-1,n} \qquad i=2,3,\ldots,n \tag{8}$$

Step 4: Compute

$$p_{i+1,j+1} = p_{i,j} - p_{i,n} a_{n-j} \qquad \begin{matrix} j=1,2,\ldots,n-2 \\ i=j+1,j+2,\ldots,n-1 \end{matrix} \tag{9}$$

The fourth step completes the evaluation of all elements of the matrix P.

It is interesting to note that the above algorithm involves only $5n^2$ multiplications more than the evaluation of a function of a diagonal matrix. Therefore, using the companion matrix approach, equation (1) involves $5n^2$ multiplications + two real matrix multiplications, that is a total of $2n^3 + 5n^2$ multiplications. On the other hand, for the diagonalization procedure equation (1) involves $4n^3 + 4n^2$ multiplications, since the matrices are complex.

The similarity transformation (2) that reduces a general matrix A to its companion form B can be easily obtained using the method of Krylov or the method of Danilevskii. Both methods are described in detail by Wilkinson[3] and Faddeeva[4]. It should be noted that the reduction of a matrix to its companion form and the computation of its eigenvalues involve, by far, less computational effort than the solution of the complete eigensystem required by the diagonalization procedure.

From the above observations it is clear that the companion matrix approach suggested here is substantially more efficient than the standard diagonalization procedure. Furthermore, since the number of operations is greatly reduced, the rounding errors involved are less severe.

REFERENCES

1. Kaufman, I., "Evaluation of an Analytical Function of a Companion Matrix with Distinct Eigenvalues", Proc. IEEE (Letters), Vol. 57, pp. 1180-1181, June 1969.
2. Kaufman, I., "A Note on the Evaluation of an Analytical Function of a Companion Matrix with Distinct Eigenvalues", to be published in the Proc. IEEE (Letters).
3. Wilkinson, J.H., The Algebraic Eigenvalue Problem. London: Oxford University Press, 1967, pp. 369-409.
4. Faddeeva, V.N., Computational Methods of Linear Algebra. New York: Dover, 1959, pp. 149-176.

ON A HOMOMORPHISM OF THE DE BRUIJN GRAPH AND ITS APPLICATIONS TO THE DESIGN OF FEEDBACK SHIFT REGISTERS*

ABRAHAM LEMPEL,† MEMBER, IEEE
Sperry Rand Research Center, Sudbury, Massachusetts

ABSTRACT

A homomorphism of the de Bruijn graph which maps a graph of order n onto one of order $n-1$ and its applications to the design of nonsingular feedback shift registers are discussed. The properties preserved under this mapping suggest a new design technique whose main advantage is due to the fact that the problem of designing a desired n-stage shift register may be reduced to a problem of order $n-1$ or less. Among the results obtained is a recursive formula for a feedback function which generates a cycle of maximum length.

1. INTRODUCTION

This paper discusses a homomorphism of the de Bruijn graph and its applications to the design of nonsingular, binary feedback shift registers (FSR's). A binary, n-stage FSR consists of n storage elements (stages) and a feedback logic producing a Boolean function $f(x_1,\ldots,x_n)$, where $x_i \in B=\{0,1\}$ indicates the binary content of stage i. When a shift pulse is applied, the state $x=(x_1,\ldots,x_n)$ of the FSR is succeeded by the state $y=(y_1,\ldots,y_n)$, where

$$y_i = x_{i+1}, \quad i = 1,\ldots,n-1 \quad \text{and} \quad y_n = f(x) \tag{1}$$

The feedback function f induces a (next-state) mapping $F:B^n \rightarrow B^n$ determined by (1). An FSR and its feedback function f are said to be *nonsingular* if the induced mapping F is one-to-one, i.e., if $xF=yF$ implies $x=y$. Evidently, an FSR is nonsingular if its state graph consists of branchless cycles.

The superposition of the state graphs for all possible n-stage FSR's is often referred to as the de Bruijn graph of order n, because of N. G. de Bruijn's early paper [1].

The problem of designing nonsingular FSR's with specified cycle lengths has received considerable attention in the literature [2-8] and most of the known theory is now available in the form of a book [8]. One of the well known results is the existence of an n-stage FSR with a cycle of length k, for any $n \geq 1$ and any $1 \leq k \leq 2^n$. This result has recently been generalized [9] also for the non-binary case. Of special interest is the problem of

* This research was carried out while the author was with the University of Southern California, Los Angeles, and was supported in part by the Army Research Office (Durham) under Contract No. DA-AROD-D-31-124G930.

† The author is with the Sperry Rand Research Center, Sudbury, Massachusetts, on leave of absence from the Technion-Israel Institute of Technology, Haifa, Israel.

finding simple methods of designing maximum length ($k = 2^n$) cycles. Although very little is known about feedback functions which generate such cycles, fairly simple methods of constructing the corresponding binary sequences are available [8, 10-12].

This paper proposes a new approach to the design of nonsingular FSR's based on the results derived in Sec. 3. This section, after some basic concepts and theorems are introduced in Sec. 2, is concerned with a homomorphism of the de Bruijn graph which maps a graph of order n onto one of order $n-1$. The properties preserved under this mapping give rise to a variety of applications, a few of which are discussed in Sec. 4. The main advantage of the proposed approach is due to the fact that, in many cases, the problem of designing a desired n-stage FSR may be reduced to a problem of order $n-1$, or less.

Among the results obtained in Sec. 4, is a recursive formula for a feedback function which generates a maximum length cycle. To the best knowledge of the author, no such formula has yet been published.

2. BASIC CONCEPTS AND THEOREMS

Consider the set of binary n-tuples B^n, formed by the n-th cartesian power of $B=\{0,1\}$. In B^n we define a relation $\rightarrow$ called the shift-relation on B^n by:

$$x \rightarrow y \text{ iff } (x_2,\ldots,x_n) = (y_1,\ldots,y_{n-1}) \tag{2}$$

The expression $x \rightarrow y$ should be read: "x shifts into y", or "y is a successor of x". For a given $x \epsilon B^n$, there are exactly two successors of x, namely, $(x_2,x_3,\ldots,x_n,0)$ and $(x_2,x_3,\ldots,x_n,1)$. Similarly, there are exactly two predecessors of x in B^n: $(0,x_1,\ldots,x_{n-1})$ and $(1,x_1,\ldots,x_{n-1})$.

The n-th order de Bruijn graph G_n is a directed graph with 2^n vertices, labelled by the elements of B^n. The verticies x and y of G_n, $x, y \epsilon B^n$, are joined by an arc $\langle x,y \rangle$, directed from x to y, iff $x \rightarrow y$.

Now let $x=(x_1,x_2,\ldots,x_n) \epsilon B^n$. We define $\hat{x}$, the <u>conjugate</u> of x, and $\bar{x}$, the <u>dual</u> of x, by $\hat{x}=(\bar{x}_1,x_2,\ldots,x_n)$ and $\bar{x}=(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n)$ where $\bar{x}_i$ denotes the Boolean complement of x_i.

Obviously,

$$\hat{\bar{x}} = \bar{\hat{x}} \tag{3}$$

and

$$x \rightarrow y \text{ iff } \hat{x} \rightarrow y \tag{4}$$

It has been observed elsewhere [5], that the one-to-one correspondence $x \leftrightarrow \bar{x}$ is the only nontrivial automorphism of the de Bruijn graph, that is,

$$x \to y \text{ iff } \bar{x} \to \bar{y} \tag{5}$$

A cycle of length k (k-cycle) in G_n is a closed sequence (ring sequence) of k distinct vertices $[x^{(1)},x^{(2)},\ldots,x^{(k)}]$ such that $x^{(k)} \to x^{(1)}$ and $x^{(i)} \to x^{(i+1)}$, $i=1,\ldots,k-1$. A convenient method of representing such a k-cycle is by means of a ring sequence of k binary digits $[c_1,c_2,\ldots,c_k]$, where $c_i = x_1^{(i)}$ is the first component of the n-tuple $x^{(i)}$, $i=1,\ldots,k$. By the shift relation between consecutive vertices in a cycle, it is clear that vertex $x^{(i)}$ is represented in the binary ring sequence by the n consecutive digits which begin with c_i. For instance, the 8-cycle $[(000),(001),(010),(101),(011),(111),(110),(100)]$ in G_3, which is a maximum length cycle for a 3-stage FSR, may simply be represented by the ring sequence $[00010111]$.

Let $C=[c_1,c_2,\ldots,c_k]$ be a k-cycle in G_n and let $x^{(i)},i=1,\ldots,k$, be the vertices of C. It follows from (5) that $\bar{C}=[\bar{c}_1,\bar{c}_2,\ldots,\bar{c}_k]$ is also a k-cycle in G_n, with vertices $\bar{x}^{(i)},i=1,\ldots,k$. The cycles C and $\bar{C}$ are said to be dual to each other. A cycle C is said to be self-dual if the ring sequence $[\bar{c}_1,\ldots,\bar{c}_k]$ is a cyclic shift of $[c_1,\ldots,c_k]$, i.e., if C and $\bar{C}$ represent the same cycle. A cycle C will be referred to as primitive if it is (vertex-) disjoint from its dual $\bar{C}$. A cycle C in G_n is called reducible if there exists a cycle C' in G_n such that the vertices contained in C' form a proper subset of those contained in C. It may easily be verified that a cycle C is reducible iff it contains at least one pair of conjugate vertices x and $\hat{x}$.

A factor of G_n is a partial graph of G_n formed by a set of disjoint cycles which, together, include all the vertices of G_n. Evidently, the state graph of every nonsingular n-stage FSR is a factor of G_n and every factor of G_n is the state graph of some nonsingular FSR. We shall use the notation (B^n,F) for the factor of G_n which is generated by the nonsingular function f.

The following are some well-known theorems which we shall need in the sequel.

Theorem 1: [5,] An FSR is nonsingular iff its feedback function $f(x)$ is of the form

$$f(x_1,x_2,\ldots,x_n) = x_1 \oplus f_0(x_2,\ldots,x_n) \tag{6}$$

where $f_0(x_2,\ldots,x_n)=f(0,x_2,\ldots,x_n)$ is an arbitrary Boolean function in the $n-1$ variables $x_2,\ldots,x_n$.

By (6), the nonsingularity of f also implies the nonsingularity of its complement $\bar{f}$.

Two cycles C_1 and C_2 are said to be adjacent if they are disjoint and there exists a vertex x in C_1 such that its conjugate $\hat{x}$ belongs to C_2.

Theorem 2: [7,8] A reducible cycle C is split into two adjacent cycles when the successors of a conjugate pair x and $\hat{x}$ in C are interchanged. Two adjacent cycles C_1 and C_2, with x in C_1 and $\hat{x}$ in C_2, are joined into a single cycle when the successors of x and $\hat{x}$ are interchanged.

We conclude this section by stating the change in the feedback function which corresponds to a split of a reducible cycle or to a join of two adjacent cycles. Let $f(x_1,\ldots,x_n)$ be the feedback function of a nonsingular FSR and let

$$h(x_1,\ldots,x_n) = f(x_1,\ldots,x_n) \oplus x_2^{a_2} \cdot x_3^{a_3} \ldots x_n^{a_n} \tag{7}$$

where $(a_2,a_3,\ldots,a_n)\epsilon B^{n-1}$ and x_i^1 and x_i^0 denote x_i and $\bar{x}_i$, respectively. The function h differs from f only for the conjugate n-tuples $a=(a_1,a_2,\ldots,a_n)$ and $\hat{a}=(\bar{a}_1,a_2,\ldots,a_n)$ and, hence:

Theorem 3: [5,8] The factors (B^n,F) and (B^n,H), corresponding to the functions f and h of (7), respectively, may be obtained from each other by interchanging the respective successors of the vertices a and $\hat{a}$.

3. THE D-MORPHISM OF THE DE BRUIJN GRAPH

Consider the set of binary n-tuples B^n. We define a mapping $D:B^n \rightarrow B^{n-1}$, $n \geq 2$, as follows:

For $a=(a_1,\ldots,a_n)\epsilon B^n$ and $\alpha=(\alpha_1,\ldots,\alpha_{n-1})\epsilon B^{n-1}$, $aD=\alpha$ iff

$$\alpha_i = a_i \oplus a_{i+1} \quad , \quad i=1,\ldots,n-1 \tag{8}$$

Given $\alpha\epsilon B^{n-1}$ and $a_1\epsilon B$, the unique solution of (8) for the remaining $n-1$ a_i's is

$$a_i = a_1 \oplus s_i \quad , \quad i=2,\ldots,n \tag{9}$$

where

$$s_i = \alpha_1 \oplus \alpha_2 \oplus \ldots \oplus \alpha_{i-1} \quad , \quad i=2,\ldots,n \tag{10}$$

Hence, we deduce:

Lemma 1: The mapping D is onto, i.e., every element in B^{n-1} is an image under D of some element in B^n.

Lemma 2: For $x, y\epsilon B^n$, $xD=yD$ iff $x=y$ or $x=\bar{y}$.

Thus, D is a mapping of B^n onto B^{n-1} under which every element in B^{n-1} is the image of a pair of dual elements in B^n. Furthermore, D preserves the shift relation, because, if $a, b\epsilon B^n$ and $a \rightarrow b$ then $b_i = a_{i+1}$,

$i=1,\ldots,n-1$, $\beta_i = b_i \oplus b_{i+1} = a_{i+1} \oplus a_{i+2} = \alpha_{i+1}$, $i=1,\ldots,n-2$, and $\alpha = aD \rightarrow bD = \beta$.

Therefore, we have:

Theorem 4: The mapping D is a homomorphism of G_n onto G_{n-1} .

If H is a subgraph of G_n and H^* is its homomorphic image, under D , in G_{n-1} , we shall refer to H^* as the D-morphic image of H and use the notation $H^* = HD$. For the remaining part of this section we shall mostly be concerned with cycles and factors of G_n such that their D-morphic images in G_{n-1} are also cycles and factors, respectively.

Noting that a primitive cycle contains no pair of dual vertices we have:

Lemma 3: The D-morphic image of a primitive k-cycle is also a k-cycle. Using the binary ring sequence representation of a cycle, it is also clear that:

Lemma 4: If $C = [c_1,\ldots,c_k]$ is a primitive k-cycle in G_n then the k-cycle $CD = \Gamma = [\gamma_1,\ldots,\gamma_k]$ in G_{n-1} is given by

$$\gamma_i = c_i \oplus c_{i+1} \quad , \quad i=1,\ldots,k-1 \tag{11}$$

and

$$\gamma_k = c_k \oplus c_1 \tag{12}$$

The weight $W(C)$ of a k-cycle $C = [c_1,\ldots,c_k]$ is defined to be the number of nonzero digits among the k c_i's , i.e.,

$$W(C) = \sum_{i=1}^{k} c_i \ .$$

Theorem 5: A k-cycle Γ in G_{n-1} is the D-morphic image of a primitive k-cycle C in G_n iff $W(\Gamma)$ is even.

Proof: If C is primitive and $\Gamma = CD$ then, by Lemma 4, we have

$$\gamma_1 \oplus \gamma_2 \oplus \ldots \oplus \gamma_{k-1} \oplus \gamma_k = (c_1 \oplus c_2) \oplus (c_2 \oplus c_3) \oplus \ldots \oplus (c_{k-1} \oplus c_k) \oplus (c_k \oplus c_1)$$

$$= c_1 \oplus c_2 \oplus c_2 \oplus c_3 \oplus \ldots \oplus c_{k-1} \oplus c_k \oplus c_k \oplus c_1 = 0$$

Thus, the modulo-2 sum of the γ's is zero and, hence $W(\Gamma)$ is even . Now, let $\Gamma = [\gamma_1,\ldots,\gamma_k]$ be a k-cycle of even weight in G_{n-1} and let

$$c_i = c_1 \oplus \gamma_1 \oplus \gamma_2 \oplus \ldots \oplus \gamma_{i-1} \quad , \quad i=2,\ldots,k \tag{13}$$

where c_1 is either 0 or 1. We claim that the dual binary ring sequences $C=[c_1,c_2,\ldots,c_k]$ and $\bar{C}=[\bar{c}_1,\bar{c}_2,\ldots,\bar{c}_k]$, with $c_1\epsilon B$ and the c_i's, $i=2,\ldots,k$, determined by (13), represent the pair of primitive and dual k-cycles in G_n whose D-morphic image is Γ. To show this, we first observe that (13) is the solution of (11) for the c_i's, $i=2,\ldots,k$, in terms of c_1 and the first $k-1$ γ's. Furthermore, since $W(\Gamma)$ is even, $\gamma_k=\gamma_1\oplus\gamma_2\oplus\ldots\oplus\gamma_{k-1}$ and (13) is also consistent with (12). Secondly, since Γ is a k-cycle in G_{n-1}, the k (n-1)-tuples formed by consecutive γ_i digits are all distinct which, in view of (11) and (12), implies that the k n-tuples formed by consecutive c_i's are not only distinct, but, also that no two of these n-tuples are dual to each other.

The following is also an immediate consequence of the proof of Theorem 5:

Corollary 1: There exists a one-to-one correspondence between the k-cycles Γ of even weight in G_{n-1} and the primitive pairs of dual k-cycles C and $\bar{C}$ in G_n under which $\Gamma=C\underline{D}=\bar{C}\underline{D}$.

We proceed now to derive an analogous correspondence in which cycles of odd weight are involved.

Lemma 5: A cycle $C=[c_i,\ldots,c_k]$ in G_n is self-dual iff it is of even length $k=2p$ and

$$c_{i+p}=\bar{c}_i \quad , \quad i=1,\ldots,p \tag{14}$$

Proof: The "if" part of this lemma is obvious. For the "only if" part, we first observe that if C is a self-dual k-cycle and x is a vertex of C then also $\bar{x}$ belongs to C and k must be even. Secondly, since the ring sequence $\bar{C}=[\bar{c}_1,\ldots,\bar{c}_k]$ is a cyclic shift of the ring sequence $C=[c_1,\ldots,c_k]$, it follows that if y is the successor of x on C then $\bar{y}$ is the successor of $\bar{x}$. This, however, is possible only if the distance between a pair of dual vertices on C is the same for all such pairs. If $k=2p$, this distance must be p and, hence, condition (14) of the lemma.

Lemma 6: The D-morphic image of a self-dual 2p-cycle is a p-cycle. If $C=[c_1,\ldots,c_{2p}]$ is a self-dual 2p-cycle in G_n then the p-cycle $C\underline{D}=\Gamma=(\gamma_1,\ldots,\gamma_p)$ in G_{n-1} is given by

$$\gamma_i = c_i \oplus c_{i+1} \quad , \quad i=1,\ldots,p \tag{15}$$

This lemma is a straightforward result of Lemma 5 and needs no further evidence.

The analogue of Theorem 5 may now be stated as follows:

Theorem 6: A p-cycle Γ in G_{n-1} is the D-morphic image of a self-dual 2p-cycle C in G_n iff $W(\Gamma)$ is odd.

Proof: From (15), we have: $\gamma_1\oplus\gamma_2\oplus\ldots\oplus\gamma_p=c_1\oplus c_{p+1}$ which, with $i=1$ in (14) reduces to $\gamma_1\oplus\gamma_2\oplus\ldots\oplus\gamma_p=1$. This proves the necessity part

of the theorem. As for the other part, let $\Gamma=[\gamma_1,\ldots,\gamma_p]$ be a p-cycle of odd weight in G_{n-1} and let

$$c_i = \gamma_1 \oplus \gamma_2 \oplus \ldots \oplus \gamma_{i-1} \quad , \quad i=2,\ldots,p \tag{16}$$

It may readily be verified, by using arguments similar to those used in the proof of Theorem 5, that the binary ring sequence $C=[0,c_2,\ldots,c_p,1,\bar{c}_2,\ldots\bar{c}_p]$, with the c_i's, $i=2,\ldots,p$, determined by (16), represents the self-dual 2p-cycle in G_n whose D-morphic image is the p-cycle Γ.

To complete the anology, we have:

Corollary 2: There exists a one-to-one correspondence between the k-cycles Γ of odd weight in G_{n-1} and the self-dual 2k-cycles C in G_n under which $\Gamma=C\underline{D}$.

The correspondences established above for individual cycles, may easily be extended to a correspondence between factors. It is clear that if a set of cycles form a factor of G_n then the set of their duals is also a factor of G_n. Two factors consisting of dual cycle sets are called dual factors and a factor is said to be self-dual if along with every cycle C in this factor also $\bar{C}$ belongs to this factor. Obviously, a factor is self-dual iff it consists of self-dual cycles and/or primitive pairs of dual cycles. From the preceding results it is also perfectly clear that:

Theorem 7: There exists a one-to-one correspondence between the factors (B^{n-1},Φ) of G_{n-1} and the self-dual factors (B^n,F) of G_n under which

$$(B^{n-1},\Phi) = (B^n,F)\underline{D} \tag{17}$$

More explicitly, relation (17) means that $B^{n-1}=B^nD$ and that for every $x\epsilon B^n,(xF)D=(xD)\Phi$. Now, if the mappings F and Φ are induced by the functions f and φ, respectively, and $(\zeta_1,\ldots,\zeta_{n-1})=(x_1,\ldots,x_n)D$, then:

$$(xF)D = (x_2,\ldots,x_n,f(x))D = (\zeta_2,\ldots,\zeta_{n-1},x_n \oplus f(x)) \tag{18}$$

$$(xD)\Phi = (\zeta_1,\ldots,\zeta_{n-1})\Phi = (\zeta_2,\ldots,\zeta_{n-1},\varphi(\zeta)) \tag{19}$$

Thus, (B^{n-1},Φ) is the D-morphic image of (B^n,F) iff $\varphi(\zeta)=x_n\oplus f(x)$, or, since $\zeta_i=x_i\oplus x_{i+1}, i=1,\ldots,n-1$,

$$f(x_1,x_2,\ldots,x_n) = x_n \oplus \varphi[(x_1 \oplus x_2),(x_2 \oplus x_3),\ldots,(x_{n-1} \oplus x_n)] \tag{20}$$

It should be noted that this expression for f in terms of φ holds not only in the nonsingular case, but also for any pair of mappings $F:B^n\rightarrow B^n$ and $\Phi:B^{n-1}\rightarrow B^{n-1}$ such that $FD=D\Phi$. In this paper, however, we are interested only in the nonsingular case and no attempt to reach beyond this case will be made.

4. SOME APPLICATIONS

The results derived in the previous section provide a basis for a new approach to the design of various types of nonsingular FSR's. Two of the possible applications are presented below.

4.1 Cycles of Maximum Length

Suppose $\Gamma=[\gamma_1,\ldots,\gamma_m]$ is an m-cycle, $m=2^{n-1}$, in G_{n-1}. It is well known that for $s\geq 2$, all the 2^s-cycles of G_s are of even weight and, therefore, by Theorem 5, Γ is the D-morphic image of a dual pair of primitive m-cycles C and $\bar{C}$ in G_n, $n>2$.

Given Γ, this pair of cycles may readily be obtained from (13) by putting $c_1=0$ and using the recursive relation

$$c_{i+1} = c_i \oplus \gamma_i \quad , \quad i=1,\ldots,m-1 \tag{21}$$

The cycles C and $\bar{C}$ are, of course, adjacent and if x and $\hat{x}$ are a pair of conjugate vertices such that x is on C and $\hat{x}$ is on $\bar{C}$ then, according to Theorem 3, interchanging the successors of x and $\hat{x}$ will result in a 2^n-cycle in G_n. Thus, to obtain a maximum length cycle in G_n, $n>2$, we have to find one $\Gamma=[\gamma_1,\ldots,\gamma_m]$ in G_{n-1}, transform it according to (21) into C and $\bar{C}$ and interchange the successors of a pair of conjugate vertices shared by C and $\bar{C}$. This simple procedure may, of course, be started at any $k<n$ for which a 2^k-cycle in G_k is known and interated $n-k$ times to obtain a maximum length cycle in G_n.

The described transition from a 2^{n-1}-cycle in G_{n-1} to a 2^n-cycle in G_n may also be expressed in terms of a recursive formula for a feedback function which will generate a maximum length cycle of an n-stage FSR. Obviously, the m-cycle Γ and the corresponding pair of cycles C and $\bar{C}$ form a pair of corresponding factors under Theorem 7. Let (B^{n-1},φ) and (B^n,F) be the factor notations for Γ and the pair C and $\bar{C}$, respectively. Thus, if φ is the feedback function generating Γ, the function f which generates the two cycles of (B^n,F) is given by (20). To obtain the function which will join these two cycles into one maximal cycle, we need to know explicitly a pair of conjugate vertices, shared by C and $\bar{C}$. Consider the vertex $e=(e_1,\ldots,e_n)$, $e_1=0$ and $e_{i+1}=\bar{e}_i$, $i=1,\ldots,n-1$. Evidently, if e belongs to C then $\bar{e}$ belongs to $\bar{C}$ and vice versa. Since e and $\bar{e}$ are on different cycles and $e\rightarrow\bar{e}$, it follows that $\bar{e}=\hat{e}F$, where $\hat{e}$ is the conjugate of e. Thus, e and $\hat{e}$ is an explicit conjugate pair that we need and, by (7), the function h which will generate a maximum length cycle of an n-stage FSR is given by

$$h(x_1,\ldots,x_n)=x_n\oplus\varphi[(x_1\oplus x_2),(x_2\oplus x_3),\ldots,(x_{n-1}\oplus x_n)]\oplus x_2^{e_2}\cdot x_3^{e_3}\cdot\ldots\cdot x_n^{e_n} \tag{22}$$

with $e_2=1$ and $e_{i+1}=\bar{e}_i$, $i=2,\ldots,n-1$.

Examples of maximal-cycle functions h_n for $n=3, 4, 5$ obtained, recursively, according to (22) from the function $h_2(x_1,x_2)=\bar{x}_1$ which generates a 4-cycle in G_2 are the following:

$$h_3(x_1,x_2,x_3) = x_3 \oplus (\overline{x_1 \oplus x_2}) \oplus x_2\bar{x}_3 = \bar{x}_1 \oplus \bar{x}_2 x_3$$

$$h_4(x_1,x_2,x_3,x_4) = x_4 \oplus (\bar{x}_1 \oplus x_2) \oplus (\bar{x}_2 \oplus x_3)(x_3 \oplus x_4) \oplus x_2\bar{x}_3 x_4$$

$$= \bar{x}_1 \oplus x_2 \oplus x_3 \oplus \bar{x}_2 x_3 \bar{x}_4$$

$$h_5 = x_5 \oplus (\bar{x}_1 \oplus x_2) \oplus (x_2 \oplus x_3) \oplus (x_3 \oplus x_4) \oplus (\bar{x}_2 \oplus x_3)(x_3 \oplus x_4)(\bar{x}_4 \oplus x_5) \oplus x_2\bar{x}_3 x_4 \bar{x}_5$$

$$= \bar{x}_1 \oplus x_2 x_3 \oplus x_4 x_5 \oplus \bar{x}_2 x_4 \oplus \bar{x}_3 x_5 \oplus \bar{x}_2 x_3 \bar{x}_4 x_5$$

where $h_5 = h_5(x_1,x_2,x_3,x_4,x_5)$.

4.2 Joining and Splitting of Cycles

The results of Sec. 3 may also be applied to simplify, in certain cases the method of designing nonsingular FSR's proposed by Yoeli [7].

It is easily observed that if a and $\hat{a}$ are a pair of conjugate vertices in G_n then aD and $\hat{a}D$ are conjugates in G_{n-1} . Also, if $\alpha = aD$ and $\beta = bD$ are conjugates in G_{n-1} then either the pairs (a,b) and $(\bar{a},\bar{b})$ or the pairs $(a,\bar{b})$ and $(\bar{a},b)$ are conjugate pairs in G_n . This observation has the following implications on the adjacency relationships between a pair of cycles Γ_1 and Γ_2 in G_{n-1} and the corresponding cycles C_1, C_2, $\bar{C}_1$ and $\bar{C}_2$ in G_n :

Case 1: The weight of both Γ_1 and Γ_2 is odd.

In this case C_i and $\bar{C}_i$ represent the same self-dual cycle in G_n and the cycles Γ_1 and Γ_2 are adjacent iff C_1 and C_2 are adjacent.

Case 2: The weight of Γ_1 is odd and that of Γ_2 is even.

In this case C_1 is self- dual, C_2 and $\bar{C}_2$ are primitive and Γ_1 and Γ_2 are adjacent iff C_1 is adjacent to both C_2 and $\bar{C}_2$.

Case 3: The weight of both Γ_1 and Γ_2 is even.

This time all four cycles C_1, $\bar{C}_1$, C_2 and $\bar{C}_2$ are primitive and Γ_1 and Γ_2 are adjacent iff either (C_1,C_2) and $(\bar{C}_1,\bar{C}_2)$ or $(C_1,\bar{C}_2)$ and $(\bar{C}_1,C_2)$ are pairs of adjacent cycles.

A similar correspondence may also be established between reducible cycles in G_{n-1} and G_n .

The above discussion suggests that the technique of joining and splitting of known cycles to obtain new cycles of specified length in G_n may, in many cases, be performed on a "cycle transition graph" [7] of order $n-1$, rather than n .

ACKNOWLEDGEMENT

The author is grateful to Professor S. W. Golomb and Professor L. R. Welch of the Department of Electrical Engineering, University of Southern California, for many helpful discussions.

REFERENCES

1. N. G. de Bruijn, "A Combinatorial Problem," Nederl. Akad. Wetensch. Proc., Vol. 49 (1946), pp. 758-764.

2. B. Elspas, "Theory of Autonomous Linear Sequential Networks," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 45-60, March 1959.

3. S. W. Golomb, L. R. Welch and R. M. Goldstein, "Cycles from Nonlinear Shift Registers," Jet Prop. Lab., California Inst. Tech., Pasadena, Prog. Rept. No. 20-389, August 1959.

4. W. W. Peterson, "Error-Correcting Codes," Technology Press of MIT, Cambridge, Mass., 1961.

5. M. Yoeli, "Nonlinear Feedback Shift Registers," IBM Dev. Lab., Poughkeepsie, N. Y., Tech. Rept. TR00.809, September 1961.

6. P. R. Bryant, F. G. Heath and R. D. Killick, "Counting with Feedback Shift Registers by Means of a Jump Technique," IRE Trans. on Electronic Computers (Correspondence), Vol. EC-11, pp. 285-286, April 1962.

7. M. Yoeli, "Counting with Nonlinear Binary Feedback Shift Registers," IEEE Trans. on Electronic Computers, Vol. EC-12, pp. 357-361, August 1963.

8. S. W. Golomb, "Shift Register Sequences," Holden-Day, Inc., San Francisco, California, 1967.

9. A. Lempel, "M-ary Closed Sequences," to appear in Journal of Combinatorial Theory.

10. L. R. Ford, Jr., "A Cyclic Arrangement of m-Tuples," Rept. No. P-1071, Rand Corp., Santa Monica, California, April 1957.

11. C. Eldert, H. J. Gray, Jr., H. M. Gurk and M. Rubinoff, "Shifting Counters," Trans. AIEE (Commun. and Electronics, No. 35), Vol. 77, pp. 70-74, March 1958.

12. E. B. Leach, "Regular Sequences and Frequency Distributions," Proc. Am. Math. Soc., Vol. 11, August 1960.

STATE ASSIGNMENTS FOR ASYNCHRONOUS SEQUENTIAL MACHINES*

by

Chung-Jen Tan**

IBM Watson Research Center
Yorktown Heights, New York

ABSTRACT

In this paper an iterative state assignment procedure for asynchronous machines is presented. Machines realized by this assignment scheme will result in circuits having a small amount of gate-inputs and operating in single transition time. It is also shown that for a flow table having m_I input columns, m input variables and d stable states the number of gate-inputs required for a single transition time realization is bounded by $d(m_I + m + 1)$.

*This work is based on the author's Doctor of Engineering Science dissertation completed in April 1969 for the Department of Electrical Engineering at Columbia University, New York and was supported by National Science Foundation grant NSF GK 1341.

**Formerly with IBM Corp., Systems Development Division, Kingston, N. Y.

STATE ASSIGNMENTS FOR ASYNCHRONOUS SEQUENTIAL MACHINES

CHUNG-JEN TAN
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

I INTRODUCTION

In a previous paper [1] it has been shown that partition pairs can be used to obtain state assignments with reduced variable dependencies for asynchronous machines. However, this approach can become quite cumbersome for asynchronous and synchronous machines. This is particularly true for the asynchronous case because critical races among state variables must be avoided in choosing a valid state assignment.

In this paper a state assignment scheme for asynchronous machines will be presented. Machines realized by this assignment scheme will result in circuits having a small amount of gate-inputs and operating in single transition time. This procedure is computationally simple in comparison with the partition pair approach. It also requires less computation than, say, finding a minimal state variable assignment by Tracey's procedure [2]. We will also establish through a constructive assignment procedure an upper bound on the gate-inputs required for a given asynchronous machine. Before we discuss the state assignment problem in detail we will first present some relevant information on asynchronous sequential machines.

An asynchronous machine is characterized by a flow table, (Fig. 1) in which each column represents an input state and each row represents an internal state. Entries of the table indicate the next states and outputs of the machine. A state i is said to be <u>stable</u> if for some input I the next state entry of i is also i, i.e. $N(i,I) = i$. This is denoted by a circled entry in the flow table. In this paper we will restrict our attention to <u>normal fundamental mode</u> machines [3]. When a machine M is realized by a sequential circuit C_M it is necessary to associate each state of M with a corresponding state of C_M. An internal state of C_M is denoted as s_i and represented internally by a binary vector y_i whose j-th component is the j-th state variable y_j.

Following are some basic definitions of asynchronous sequential machines:

<u>Definition 1</u>. If $N(s_i,\underline{x}_k) = s_j$, then the change of values of state variables from the internal state code associated with state s_i to the code

associated with state s_j is said to be a transition from s_i to s_j.

Definition 2. An internal state s_r is said to be an intermediate state between states s_i and s_j if the state variables may assume the value associated with the internal state s_r during the transition from s_i to s_j.

Definition 3. If $N(s_i, \underline{x}_k) = s_j$ and $N(s_m, \underline{x}_k) = s_n$, where $j \neq n$, and both transitions share a common intermediate state s_c then the circuit C_M is said to have a critical race.

Definition 4. A state assignment is called a single transition time (STT) assignment when all state variables that are required to change during a transition are allowed to change simultaneously without critical races. Further, if only a single coding is associated with each state, it is called a uni-code single transition time (USTT) assignment.

USTT- assignments were first derived by Liu [4] and later extended by Tracey [2]. Their assignment schemes were concerned with minimizing the number of state variables required. However, minimal variable assignment does not necessarily yield economical circuit realization.

For example, consider the Machine A in Fig. 1 below.

x_1x_2

	I_1 00	I_2 01	I_3 11	I_4 10	Z	Assignment 1 y_1	y_2	y_3	y_4	Assignment 2 y_5	y_6	y_7	y_8	y_9	y_{10}
1	(1)	(1)	5	(1)	1	0	1	0	1	0	0	0	0	0	0
2	1	(2)	5	4	0	0	0	1	0	0	0	1	1	0	0
3	(3)	1	5	(3)	0	1	0	0	0	0	0	0	0	1	1
4	3	(4)	5	(4)	0	1	0	1	1	0	1	1	0	1	0
5	(5)	2	(5)	1	0	1	1	1	0	1	0	0	1	0	0

Figure 1 - Machine A.

Assignment 1 is a minimal variable assignment. Their respective next state equations are:

$$
\begin{aligned}
Y_1 &= \bar{x}_1\bar{x}_2y_1 + x_2y_3y_4 + x_1x_2 + x_1\bar{y}_2 \\
Y_2 &= \bar{x}_1x_2y_2 + x_1\bar{x}_2y_1 + x_2y_3\,x_1x_2 + x_1y_2 \\
Y_3 &= \bar{x}_1x_2y_1y_2 + x_2\bar{y}_3 + x_1x_2 + x_1y_3\bar{y}_2 \\
Y_4 &= \bar{x}_1x_2y_1 + x_1\bar{x}_2y_3 + x_1x_2y_4 + x_1x_2y_2 + x_1\bar{x}_2y_3
\end{aligned}
\qquad (1)
$$

$$
\begin{aligned}
Y_5 &= \bar{x}_1\bar{x}_2y_5 + x_1\bar{x}_2 \\
Y_6 &= \bar{x}_1x_2y_6 + x_1\bar{x}_2y_7 \\
Y_7 &= \bar{x}_1\bar{x}_2y_6 + \bar{x}_1x_2y_8 + x_1\bar{x}_2y_7 \\
Y_8 &= \bar{x}_1\bar{x}_2y_5 + \bar{x}_1x_2y_8 + x_1\bar{x}_2 \\
Y_9 &= \bar{x}_1\bar{x}_2y_9 + x_1\bar{x}_2y_6 + x_1x_2y_7 + x_1\bar{x}_2y_{10} \\
Y_{10} &= \bar{x}_1x_2y_9 + x_1x_2y_{10}
\end{aligned}
\qquad (2)
$$

Counting shared product terms in the above realizations assignment 1 yields a circuit having 65 gate-inputs while assignment 2 yields a circuit having only 36 gate-inputs. In fact we will show in the next section that the first realization above has more gate-inputs than that yielded by an easily derived upper bound for machine A.

II. UPPER BOUND ON THE GATE-INPUTS REQUIRED FOR USTT-REALIZATIONS

In this section we will first present a simple assignment scheme which will result in a Liu-assignment [4] for any given machine. A simple formula for calculating the upper bound of gate-inputs required for USTT-realizations will then be derived from this assignment scheme. Notice that if a state variable $y_i = 1$ for state b then the next value for the i-th state variable, Y_i, should be 1 whenever we have a transition a $\xrightarrow{e}$ b under some input column I_e. Hence $Y_i = 1$ during all intermediate states between a and b under input I_e. Therefore, the subspace spanned by the total states(I_e,a) and (I_e,b) in the total state space must be covered by a product term in the equation for Y_i [5]. Furthermore, if c $\xrightarrow{e}$ d and $y_i = 0$ for state d then the product term covering the subspace spanned by (I_e,a) and (I_e,b) must not contain any intermediate states between states c and d under input I_e.

Definition 5. A transition a $\xrightarrow{e}$ b will be called a 1-transition with respect to variable y_i if $y_i = 1$ for b and 0-transition if $y_i = 0$ for b.

It follows from above discussions that the space spanned by each product term, in a next state equation for some variable y_i, must not contain any intermediate states of some states involved in a 0-transition and the space spanned by each 1-transition must be covered by at least one product term in the equation for Y_i.

Definition 6. A product term in the next state equation for a variable y_i is said to be a simple product term if it contains only one state variable.

The assignment procedure proposed in this section will make the next state equations contain only simple product terms for all of the variables.

Let the set of states $D_{i,q} = \{s \mid N(s,I_i) = q\}$ be called a destination set [5] under input I_i, and let D_{I_i} denote the set of destination sets under I_i.

Consider the flow table and the accompanying assignment in Figure 2.

X	I_1	I_2	y_1	y_2	y_3	y_4	y_5	y_6
	0	1						
1	(1)	2	1	0	0	1	0	0
2	3	(2)	0	1	0	1	0	0
3	(3)	4	0	1	0	0	1	0
4	5	(4)	0	0	1	0	1	0
5	(5)	6	0	0	1	0	0	1
6	1	(6)	1	0	0	0	0	1

Figure 2.

We have $D_{1,1} = \{6,1\}$, $D_{1,3} = \{2,3\}$, and $D_{1,5} = \{4,5\}$. The set D_{I_1} will be denoted as $D_{I_1} = (6\textcircled{1}, 2\textcircled{3}, 4\textcircled{5})$. Similarly, $D_{I_2} = (1\textcircled{2}, 3\textcircled{4}, 5\textcircled{6})$. Variable $y_1 = 1$ for states in $D_{1,1}$ and $y_1 = 0$ for states in $D_{1,3}$ and $D_{1,5}$. Hence the product term for transitions under I_1 in the equation for Y_1 will be $\bar{x}y_1$. Transition $5 \overset{x}{\rightarrow} 6$ is the only 1-transition with respect to y_1 in I_2. The variable separating $5 \overset{x}{\rightarrow} 6$ from all other 0-transitions in I_2 is y_6. Therefore, in the equation for Y_1 the product term covering $5 \overset{x}{\rightarrow} 6$ will be xy_6. Thus we have $Y_1 = \bar{x}y_1 + xy_6$.

Lemma 1: For a given machine, if there is a state variable y_j^i for each destination set D_j^i such that $y_j^i = 1$ for $s \in D_j^i$ and $y_j^i = 0$ for $s \notin D_j^i$ then all the next state equations can be expressed in terms of simple product terms and the circuit will have no critical races.

Proof.

Consider a variable y_j^i where $y_j^i = 1$ for states contained in D_j^i only. The transitions under some input I_e will consist of 1-transitions and 0-transitions with respect to y_j^i. If D_t^e consists of states involved in 1-transitions then the product term $x_e y_t^e$ covers the space spanned by all 1-transitions involving states in D_t^e and contains no intermediate states for 0-transitions under I_e. Therefore, for each destination set containing states involved in 1-transitions there is a unique product term, in the equations for Y_j^i, which separates these transitions from all other 0-transitions. Hence there are no critical races. Furthermore, it is obvious that every 1-transition, with respect to y_j^i, can be covered by a simple product term. The above results are true for every state variable. Thus all the next state equations can be made to consist of simple product terms.

For a given table, we will assign a unique variable to each distinct destination set so that each next state equation will consist of simple product terms. For the flow table in Fig. 2 we have

$D_{I_1} = (6\textcircled{1}, 2\textcircled{3}, 4\textcircled{5})$, $D_{I_2} = (1\textcircled{2}, 3\textcircled{4}, 5\textcircled{6})$.

Let's order the destination sets such that

$D_1^1 = \{1,6\}$, $D_2^1 = \{2,3\}$, $D_3^1 = \{4,5\}$, $D_4^2 = \{1,2\}$

$D_5^2 = \{3,4\}$, $D_6^2 = \{5,6\}$.

For each destination set D_j^i we associate it with a state variable y_j^i such that $y_j^i = 0$ for states not contained in D_j^i. It is easy to see that the next state equation for each of these variables can be expressed in terms of simple terms, with each simple product term covering the unique destination set containing states involved in a 1-transition. The accompanying assignment for the flow table in Fig. 2 is exactly such an assignment. The set of next state equations are:

$$
\begin{aligned}
Y_1^1 &= \bar{x}y_1^1 + xy_6^2 \\
Y_2^1 &= \bar{x}y_2^1 + xy_4^2 \\
Y_3^1 &= \bar{x}y_3^1 + xy_5^2 \\
Y_4^2 &= \bar{x}y_1^1 + xy_4^2 \\
Y_5^2 &= \bar{x}y_2^1 + xy_5^2 \\
Y_6^2 &= \bar{x}y_3^1 + xy_6^2
\end{aligned}
\tag{3}
$$

For a given table if there exists an input column I_e such that all the specified next state entries are identical, such as I_3 in Fig. 1, then the simple product term for that column will be $\underline{x}_e$ if these are all 1-transitions with respect to the variable in question. Hence no variable is needed for the only destination set in this column. Furthermore, if there exists an input column I_e where all the specified entries are stable then in the next state equation for y_i^j, where $j \neq e$, the simple product term for states in I_e will be $\underline{x}_e y_i^j$. Therefore, if all the stable states in I_e are contained in different destination sets in other input columns, namely each state will have a unique coding, then no variable is needed for the destination sets in I_e. These columns, where no state variable is needed under our scheme will be called trivial columns. The following is a summary of the above assignment scheme.

Assignment Procedure 1

1. Order the destination sets, $D_{i,j}$, of a given flow table except those contained in trivial input columns.

2. Associate a unique variable y_t^e for each element in the set of ordered destination sets, D_t^e, such that $y_t^e = 1$ for states in D_t^e and $y_t^e = 0$ for states not in D_t^e.

3. For each variable y_j^i express the next state equation as a sum of simple product terms such that a simple product term $x_e y_t^e$ is contained in the equation if and only if states in D_t^e are involved in 1-transitions with respect to y_j^i.

Notice that in Equation 3 each product term is shared by two equations. Thus using shared logic savings in gate-inputs needed are achieved. Furthermore the assignment procedure is very easy to carry out. For many flow tables the circuit so realized will actually require less gate-inputs than that needed for some minimal variable assignments.

Procedure 1 above also implies an upper bound on the gate-inputs required for any given normal fundamental mode flow table. If a table does not have an input column whose specified entries are all stable then the number of simple product terms in the equation for Y_j^i equals the number of destination sets, $D_{e,q}$, whose stable states are contained D_j^i. It is also clear that the number of distinct state variables is equal to the number of distinct destination sets excluding those in the trivial columns. Therefore, if a flow table has d_s distinct nontrivial destination sets and d stable total states in columns having at least one unstable next state entry and t trivial columns whose specified entries are all stable, then we can obtain a USTT-realization for this table with a set of next state equations consisting of d_s state variables with $d + td_s$ distinct simple product terms. If h_i equals the number of stable states, in columns having at least one unstable entry, that are contained in the i-th destination set, D_i^e, then the equation for Y_i^e will have $h_i + t$ product terms. Since a simple product term contains m input variables and one state variable and a product term for a trivial column having only one stable state contains m input variables only, we obtain immediately the following result.

Theorem 1.

For a given flow table if

d = number of stable total states in columns having at least one unstable next state entry,

d_s = number of distinct nontrivial destination sets, D_i^e,

h_i = number of stable states, in columns having at least one unstable entry, that are contained in the i-th destination set D_i^e,

m = number of input variables

l = number of trivial input columns having at least one unstable state and only one stable state,

t = number of trivial input columns whose specified entries are all stable,

then the number of gate-inputs required for a USTT-realization is less than or equal to

$$\sum_{1 \le i \le d_s} h_i + (d + d_s - t)(m + 1) + ml. \qquad (4)$$

For machine A in Fig. 1 we have $d_s = 9$, $d = 10$, $t = 0$, $l = 1$, and $m = 2$. So the upper bound is equal to $30 + (10)(3) + 2 = 62$. This is smaller than that resulted from the minimal variable assignment shown in Eq. (1).

Corollary 1.

The number of gate-inputs required for a flow table without trivial columns is bounded by

$$\sum_{1 \le i \le d_s} h_i + d(m+1) \qquad (5)$$

Let us consider fully specified tables without trivial input columns. Let's assume that all destination sets for a given table are all distinct. The total number of product terms in the equations for variables associated with destination sets in some column I_e will be identical to the total number of stable total states in the table. This fact is true for all input columns. Hence, if m_I = number of input columns in the table then $\sum_{i=1}^{d} h_i = m_I d$. Thus we have the following result.

Corollary 2.

Given a flow table the number of gate-inputs required for a USTT-realization is less than or equal to

$$d(m_I + m + 1). \qquad (6)$$

The bound derived above increases as a function of the number of stable states of the flow table. It is interesting to compare this bound with that obtained by the 1-hot assignment [5]. If m is the number of input variables, r the number of states, and u the number of unstable states, then an upper bound on the number of gate-inputs required for a 2-stage implementation of the next state equations for a 1-hot assignment is $2r+u(m+3)$[5]. Notice that this bound tends to increase as a function of the number of unstable states of the table. Therefore, for flow tables having less number of stable states than unstable states the bound derived in this section will in general be better than that derived from the 1-hot code. For example, the 1-hot code for the flow table in Fig. 1 will require 65 gate inputs.

It should also be pointed out that the 1-hot code requires 2-step transition while the assignment procedure discussed in this section will result in circuits operating in single transition time. There does not exist, at present, any nontrivial bound for other modes of circuit operation.

III GENERAL STATE ASSIGNMENT PROCEDURES

We will be concerned, in this section, with algorithms for state assignments. A specific procedure resulting in a USTT-assignment with Liu-type of state variables will also be presented. The procedure is a refinement pf procedure 1 above. It yields an assignment for a given table such that most, but not all, of the next state equations contain simple product terms that are shared by two or more variables. Let us consider first the general assignment problem.

Definition 7. [5] Let U and V be disjoint subsets of the states of a machine. The unordered pair (U,V) is referred to as a partial state dichotomy (or simply, dichotomy) of the machine. A state variable y_i is said to cover a dichotomy if $y_i = 0$ for all $s_j \in U$ and $y_i = 1$ for all $s_k \in V$, or vice versa. If any column of a flow table contains the transitions $i \rightarrow j$ and $k \rightarrow m$, we say that the dichotomy (ij,km) is associated with the transitions. If $i = j$, we have a degenerated dichotomy (i,km); if $i = j$ and $k = m$, we have the degenerate dichotomy (i,k).

Definition 8. Two state transitions $i \rightarrow j$ and $k \rightarrow m$, in some input column I_t, are said to be distinct if states j and m are not equivalent.

Definition 9. A dichotomy (ij, km) associated with the transitions $i \rightarrow j$ and $k \rightarrow m$ is said to be relevant to a state variable y_t if $y_t = 0$ for j and $y_t = 1$ for m, or vice versa.

Tracey [2] has shown that the necessary and sufficient conditions for a USTT-assignment for a normal mode flow table are: (i) the dichotomy associated with every pair of transitions occurring in any column of the flow table should be covered by some state variable; (ii) there must be a unique coding for every state. It can be shown that in order to obtain a valid USTT-assignment it is sufficient to cover the set of all dichotomies associated with distinct state transitions [6]. Thus any state assignment such that the set of state variables cover all dichotomies associated with columnar state transitions will yield a valid USTT-realization. Procedures have been established previously for obtaining assignments with reduced variable dependencies [1]. However, these procedures as well as that established by Tracey [2] for obtaining minimal variable assignment all require extensive enumeration. Furthermore, these assignments do not always

yield economical circuit realizations. To bypass these difficulties an iterative procedure is proposed below. For a given machine a set of relatively good state variables are first chosen to yield an initial partial state assignment. Procedures are established to derive critical-race-free next state equations in the sum-of-product form for these initial set of state variables. In the process of deriving next state equations new variables will be generated. Next state equations will be derived for these newly generated variables until a sufficient number of state variables have been chosen to cover the set of all columnar dichotomies. At each step the newly generated variables are chosen according to some heuristic rule. Thus variant forms of this procedure can be established.

Theorem 2.[1,6]

Given a state assignment with variables $S_y = \{y_i\}$ where $y_0 \in S_y$ and $\sigma_0 \subseteq S_y$, $Y_0 = f_0(\underline{X},\sigma_0)$ iff all dichotomies relevant to y_0 are covered by some $y_i \in \sigma_0$.

From the above theorem it follows that once we have chosen a variable y_0 and determined the corresponding set σ_0 the set of variables $\{y_0\} \cup \sigma_0$ will constitute a partial state assignment. By repeating this process for other variables in σ_0 enough variables will eventually be chosen which cover the set of all columnar dichotomies. Hence a valid USTT-realization can be obtained.

For example, consider the Machine A in Fig. 1 repeated below together with the first assignment.

x_1x_2	I_1 00	I_2 01	I_3 11	I_4 10	Z	y_1	y_2	y_3	y_4
1	(1)	(1)	5	(1)	1	0	1	0	1
2	1	(2)	5	4	0	0	0	1	0
3	(3)	1	5	(3)	0	1	0	0	0
4	3	(4)	5	(4)	0	1	0	1	1
5	(5)	2	(5)	1	0	1	1	1	0

Figure 3 - Machine A.

Table 1 illustrates the set of dichotomies of Machine A and their relations to each state variable.

x_1x_2		y_1	y_2	y_3	y_4
00	$d_1 = (21,43)$	✓	⊬	-	⊬
	$d_2 = (21,5)$	✓	-	⊬	⊬
	$d_3 = (43,5)$	-	✓	⊬	-
01	$d_4 = (31,52)$	-	⊬	✓	⊬
	$d_5 = (31,4)$	⊬	⊬	✓	-
	$d_6 = (52,4)$	⊬	-	-	✓
11	----------	-	-	-	-
	$d_7 = (24,51)$	⊬	✓	⊬	⊬
10	$d_8 = (24,3)$	-	-	✓	⊬
	$d_9 = (51,3)$	⊬	✓	-	⊬

-: means not relevant

⊬: means relevant but not covered by y_i

✓: means relevant and covered by y_i

Table 1.

If variable y_2 is first chosen then we could let $\sigma_2 = \{y_1, y_2, y_3\}$. Consequently, $Y_2 = f_2(x_1, x_2, y_1, y_2, y_3)$. Now we can choose $\sigma_1 = \{y_1, y_2, y_3, y_4\}$ and let $Y_1 = f_1(x_1, x_2, y_1, y_2, y_3, y_4)$. Since the four variables cover the set of all columnar dichotomies we have obtained a valid assignment. Notice that for a given variable y_0 the set σ_0 is not unique. Different choice of σ_0 will yield different assignment. Consequently, different circuit realizations will arise. The assignment scheme specified in Procedure 1 in the previous section not only indicates the choice of a particular variable y_0 and the associated set σ_0 it also indicates how Y_0 is to be expressed as a function of the elements in σ_0 and the input variables. We will now show that two level sum-of-product form next state equations can also be obtained directly for any variable y_0 and any one of its associated state variable subset σ_0. This enables us to establish rules so that a particular σ_0 can be chosen in an iterative assignment scheme.

Let $a \xrightarrow{e} b$ be a 1-transition with respect to a variable y_i under input I_e, $\Delta_i^{ab,e} = \{(ab,cd)\}$ be the set of dichotomies relevant to y_i and associated with transition $a \xrightarrow{e} b$. Since $a \xrightarrow{e} b$ is a 1-transition it thus follows that $c \xrightarrow{e} d$ must be a 0-transition with respect to y_i. Therefore, in the product term covering the 1-transition $a \xrightarrow{e} b$, the space spanned by this term must not contain any states between (I_e,c) and (I_e,d). Let $\sigma_i^{ab,e} \subseteq \sigma_i$ be a subset of the set of state variables covering

elements of $\Delta_i^{ab,e}$. Since for each 0-transition $c \xrightarrow{e} d$ the dichotomy (ab,cd) is relevant to y_i, thus covered by some $y_i \varepsilon \sigma_i^{ab,e}$, variables in this set separate the 1-transition $a \xrightarrow{e} b$ from all other 0-transitions under I_e.

Definition 10. If A is a subset of the set of all states then we let $y_t^A = y_t$ if $y_t = 1$ for all the states in A and $y_t^A = \bar{y}_t$ if $y_t = 0$ for all the states in A. Otherwise undefined.

From above definition it follows that the subspace spanned by variables in $\sigma_i^{ab,e}$ and input $\underline{x}_e$, containing the subspace between (I_e,a) and (I_e,b), is represented in the next state equation for y_i by the product term

$$P_j = \underline{x}_e \prod_{\sigma_i^{ab,e}} y_t^{\{a,b\}} \tag{7}$$

If there exists an input column I_e consisting of no 0-transitions, then the 1-points in column I_e are all covered by the subspace spanned by the input vector $\underline{x}_e$ itself.

For example, in Fig. 3, the 1-transition $2 \xrightarrow{4} 4$ with respect to y_3 is associated with the set of dichotomies $\Delta_3^{24,44} = \{(24,51),(24,3)\}$. Notice from Table 1 that the first dichotomy is covered by y_2 while the latter is covered by y_3. Thus $\sigma_3^{24,4} = \{y_2,y_3\}$. Since $y_2 = 0$ for states $\{2,4\}$ and $y_3 = 1$ for states$\{2,4\}$it follows that $y_2^{\{2,4\}} = \bar{y}_2$, $y_3^{\{2.4\}} = y_3$. The following is a summary of the above discussion.

Theorem 3.

Let T_0 be the set of 1-transitions with respect to the state variable y_0, Δ_0 the set of relevant dichotomies, and $\Delta_0^{ab,e} \subseteq \Delta_0$ the subset of dichotomies associated with 1-transition $a \xrightarrow{e} b$. If σ_0 is the subset of state variables covering elements in Δ_0 and $\sigma_0^{ab,e} \subseteq \sigma_0$ covers the elements in $\Delta_0^{ab,e}$ then

$$Y_0 = f_0(\underline{x},\sigma_0) = \sum_{T_0} \underline{x}_e \prod_{\sigma_0^{ab,e}} y_i^{\{a,b\}}$$

Notice that in Assignment Procedure 1 we have implicitly applied this result.

We will now discuss a specific iterative assignment scheme. This scheme is essentially a finement of Procedure 1 presented in the previous section. Notice that a product term will be shared by two or more state variables if identical set of states, namely those contained in some destination sets, are involved in 1-transitions with respect to these variable. If variables associated with these sets are used in the assignment then shared simple product terms can be obtained in the next state equations.

Thus we will choose a subset of those variables which dichotomize some destination sets for assignment. Theorem 3 will also be applied in the procedure to obtain directly the next state equations.

Consider the choice of the initial set of variables. Notice that if a variable y_i dichotomizes both D_{I_e} and D_{I_m}, where I_e and I_m are two different input columns, then by using y_i in the assignment we will very likely have saved one variable. Furthermore, in the expression for Y_i we will have two product terms $x_e y_i$ and $x_m y_i$. If I_e and I_m are adjacent columns then the expression may be further simplified. Notice also that if a variable $y_i = 0$ for all the stable states in input I_e then there is no product term for transitions under I_e. Hence the cost for Y_i will likely be small. Finally, if a variable y_i has, in the equation for Y_i, most of its product terms appeared in some other equations then the cost for Y_i is also likely to be small. We will first find those y_i's that dichotomize more than one set of destination sets. Those variables for which only a few input columns have 1-transitions will also be found. The next state equations for these variables will be derived and denoted symbolically by equations of the form

$$Y_0 = [ab...q] = ...+ I_i\,[de...r] + ... + I_e\,[fg...s]^*, \qquad (8)$$

where $Y_0 = 1$ for states a,b,...,q ; states d,e,...,r are those involved in 1-transitions under input I_i and covered by one of the variables in the initial set; f,g,...,s are those states involved in 1-transitions under I_e but are not covered by any variable that have already been chosen. There will be one product term in Eq. (8) for each input I_i. New variables will then be chosen to cover those 1-transitions in a starred product term. The process of adding new variables terminates when every pair of elements in each set D_{I_e} are separated by at least one state variable. Notice that this is exactly the condition for a valid modified Liu assignment [4,5]. Thus for each variable y_0 we will have found the corresponding set σ_0. The next state equations for these variables can be obtained according to Theorem 3.

In the process of deriving new variables we first assign a distinct variable y_i^* to states in each starred product term $I_e[ab...q]^*$, Thus in Eq. (8) the product term $I_e[ab...q]^*$ is logically equivalent to $I_e y_i^*$. Next state equations for these variables are then derived. Sometimes a product term involving a starred variable $I_e y_i^*$, namely a variable yet to be chosen, can be replaced by a sum of simple product terms involving y_i's already chosen and a simple product term involving a different starred variable. That is

$$I_e y_i^* = I_e y_1^c + I_e y_2^c + \ldots + I_e y_k^c + I_e y_j^*$$

where y_i^c denotes a variable y_i that has already been chosen for assignment. If this is the case then we will denote it as

$$I_e[y_1^c + y_2^c + \ldots + y_k^c + y_j^*]^*.$$

Each starred variable will be given a cost measure. At each step, if we have to choose a starred variable then the one with the least cost will be chosen.

Let t_{y_i} be the number of product terms, in the equation for Y_i, that do not involve a starred variable and $t_{y_i}^*$ be the number of product terms involving a starred variable. Let k_i equal the minimum additional variables still needed to separate every pair of destination sets in D_{I_i}. Let m_s denote the number of input columns such that the choice of y_i reduces the respective number k_i by 1.

One possible cost for a given variable y_i is given by

$$C_m(y_i) = t_{y_i} + (m+2)t_{y_i}^* - (m+1)(m_s - 1)m_I \tag{9}$$

If the table is incompletely specified then the cost for a starred variable is dependent upon how the don't cares are utilized. Hence attempts should be made to find the most economical way of using the don't cares. This, fortunately, does not complicate the procedure.

The cost for the product term $I_e y_i^* = I_e[y_i^c + y_2^c + \ldots + y_k^c + y_j^*]^*$ is given by

$$\min(C_m(y_i^*),\ C_m(y_j^*) + k) \tag{10}$$

If $C_m(y_j^*) + k < C_m(y_i^*)$ then the product term $I_e y_i$ will be replaced by

$$\sum_{j=1}^{k} I_e y_j^c + I_e y_j^*.$$

Notice that the cost measure of the set of variables yet to be chosen is dependent upon those variables that have been used for assignment. Hence the cost measure should be updated after each variable has been chosen. Furthermore, in order to find the number t_{y_i} and $t_{y_i}^*$ we should derive the next state equations for each of these variables. This is not a wasteful procedure because most of these variables will eventually be chosen.

When enough variables have been chosen for each set D_{I_e}, those product terms involving a starred variable will be replaced by the set of chosen

variables. If there are nonsimple product terms, in the final set of equations, we will change each of these terms to a simple product term by introducing a new variable only if this will reduce the total cost. Variables not needed to cover any columnar dichotomies will be deleted if this will also reduce the total cost. The following is a summary of the assignment scheme.

Assignment Procedure 2.

1. Find those y_i's that dichotomize more than one set of destination sets as well as those for which only a few input columns have 1-transitions.

2. From the above set of variables find one or more initial subsets for partial state assignment.

3. Derive the next state equations for these initial variables in the form of Eq. (8).

4. Among the starred variables choose one for state assignment according to the cost measure given by Eqs. (9) and (10).

5. Repeat step 4 until every destination set in each individual column are separated from one another by some chosen variable.

6. Apply Theorem 3 to obtain the sum-of-product form next state equations.

7. Delete or add new variables if by doing so the logic equations can be simplified.

It should be noted that there exist definite procedures for finding the initial set of good variables [6]. However, Since this is rather straightforward we will not discuss it here. Furthermore, different heuristic rules can be established to accomplish this task as well as that of choosing new variables. We will illustrate this approach by obtaining an assignment for the flow table of Machine B in Fig. 4.

x_1x_2	I_1	I_2	I_3	I_4
1	(1)	2	(1)	3
2	3	(2)	1	5
3	(3)	(3)	4	(3)
4	3	2	(4)	5
5	(5)	2	6	(5)
6	1	2	(6)	(6)

Figure 4. Machine B.

The destination sets of the flow table for Machine B are:

$$D_{I_1} = (\textcircled{1}\ 6,\ 2\ \textcircled{3}\ 4, \textcircled{5}\ ).$$
$$D_{I_2} = (\ 1\ \textcircled{2}\ 456,\ \textcircled{3}\ ),$$
$$D_{I_3} = (\textcircled{1}\ 2,\ \ 3\ \textcircled{4}, 5\ \textcircled{6}\ ),$$
$$D_{I_4} = (1\ \textcircled{3}\ ,\ 24\ \textcircled{5}\ ,\ \textcircled{6}\ ).$$

It can be checked that there does not exist variables that dichotomize more than one set of destination sets. Two block state partitions that can be used to induce variables such that transitions in one or more input columns are all 0-transitions with respect to these variables are:

$$\rho_1^1 = (12346,5), \rho_2^1 = (2345,16), \rho_1^3 = (3456,12),$$

$$\rho_2^3 = (1234,56), \rho_1^4 = (12345,6).$$

However, since ρ_2^3 and ρ_2^1 together separate $D_{1,5_1}$ from $D_{1,1}$ and $D_{1,3}$, ρ_1^1 is not needed in an assignment. Similarly, ρ_2^3 and ρ_2^1 together separate $D_{4,6}$ from $D_{4,3}$ and $D_{4,5}$ hence we don't need ρ_1^4. Let the initial set of variables be induced by ρ_2^1, ρ_1^3, and ρ_2^3. We have

$$Y_1 = [16] = I_1[16] + I_3[1256]^* + I_4[6]^*$$
$$Y_2 = [12] = I_1[16] + I_2[12456]^* + I_3[12]$$
$$Y_3 = [56] = I_1[5]^* + I_3[56] + I_4[2456]^*$$

Now we have enough variables to separate every pair of destination sets in D_{I_1} and D_{I_3} respectively. We need one more variable each for D_{I_2} and D_{I_4}. So we let

$$Y_4 = [12456] = I_1[16] + I_1[5]^* + I_2[12456] + I_3 + I_4[2456]^*, \text{ and}$$

$$Y_5 = [2456] = I_1[5]^* + I_2[12456] + I_3[3456]^* + I_4[2456].$$

By substituting appropriate variables to the above set of equations we have the following next state equations:

$$
\begin{aligned}
Y_1 &= \underline{x}_1 y_1 + \underline{x}_3 y_2 + \underline{x}_3 y_3 + \underline{x}_4 y_1 y_3 \\
Y_2 &= \underline{x}_1 y_1 + \underline{x}_2 y_4 + \underline{x}_3 y_2 \\
Y_3 &= \underline{x}_1 \bar{y}_1 y_3 + \underline{x}_3 y_3 + \underline{x}_4 y_5 \\
Y_4 &= \underline{x}_1 y_1 + \underline{x}_1 \bar{y}_1 y_3 + \underline{x}_2 y_4 + \underline{x}_3 + \underline{x}_4 y_5 \\
Y_5 &= \underline{x}_1 \bar{y}_1 y_3 + \underline{x}_2 y_4 + \underline{x}_3 \bar{y}_4 \bar{y}_3 + \underline{x}_3 y_3 + \underline{x}_4 y_5
\end{aligned}
\qquad (11)
$$

The set of equations in (11) requires 49 gate-inputs. However, we note that y_4 is not needed to cover any columnar dichotomies of the flow table. Hence we can replace each product term in (11) involving y_4 by a set of different product terms. The product term $\underline{x}_2 y_4$ can be replaced by $\underline{x}_2 y_5 + \underline{x}_2 y_2$. But if y_4 is deleted the product term $\underline{x}_3$ in the equation for Y_4 can be dropped. Hence the total cost will become $49 + 1 - (5 + 2) = 43$. It should be noted that if we introduce some new variables then we can replace some of the nonsimple product terms in (11) by simple product terms. However, this will not decrease the cost in this case. The second assignment for Machine A in Fig. 1 is also obtained by this procedure.

IV CONCLUSIONS

In this paper we have presented a specific USTT-assignment procedure for normal fundamental mode asynchronous flow tables. An easily calculated upper bound on gate-inputs required for any given table is also derived from this simple procedure. A general iterative approach for obtaining state assignments is also established. This is illustrated by a specific assignment scheme. It should be pointed out that this heuristic procedure yields relatively good results even for relatively large flow tables. Furthermore, heuristic rules yet to be established by further research can be easily incorporated into this procedure. Thus, it does present a good alternative to the existing methods of state assignments.

ACKNOWLEDGMENT

The author wishes to thank Professor S. H. Unger for his guidance and encouragement during the course of this work.

REFERENCES

[1] C. J. Tan, P. R. Menon, A. D. Friedman, "Structural Simplification and Decomposition of Asynchronous Sequential Circuits," IEEE 9th Annual Symposium on Switching and Automata Theory, Schenectady, N. Y., Oct. 1968. Also in IEEE Transaction on Computers, Sept. 1969.

[2] J. H. Tracey, "Internal State Assignment for Asynchronous Sequential Machines," IEEE-TEC, V. EC-15, No. 4, pp. 551-560, Aug. 1966.

[3] E. J. McCluskey, "Introduction to Switching Circuit Theory," McGraw-Hill, 1965.

[4] C. N. Liu, "A State Variable Assignment Method for Asynchronous Sequential Machines," J. ACM, Vol. 10, pp. 209-216, April 1963.

[5] S. H. Unger, "Theory of Asynchronous Sequential Switching Circuits," John Wiley and Sons, Inc., New York, N. Y., to be published in 1969.

[6] C. J. Tan, "Synthesis of Asynchronous Sequential Switching Circuits," Dr. Eng. Sc. Dissertation, Columbia University, June 1969.

COST CONSIDERATIONS OF BASE N ADDERS USING N-VALUED LOGIC ELEMENTS

Anthony S. Wojcik
and
Gernot Metze
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

This paper considers the use of N-valued logics in the design of base N adder circuits. General formulas for the sum and carry functions of an adder are established, and relative costs of the base N adders are discussed in terms of a cost function and a break-even cost bound.

INTRODUCTION

With recent advances in fluidics and integrated circuit technology, interest in higher-order logics and multi-valued switching theory has grown sharply. As a result, several authors have investigated the theoretical properties of N-valued logics and have established a basic body of material on N-valued combinational switching theory [1-4]. Of particular interest has been the development of ternary switching theory [5-7]. The question of what use can be made of this generalized switching theory for combinational circuits is still open.

The purpose of this paper is to consider the use of N-valued logics in the design of base N adder circuits and to evaluate the relative costs of these adders.

N-VALUED LOGIC

The N-valued combinational switching theory employed in this paper is based on Post algebras [8-9]. For an N-valued system, the set of N physical values, e.g. voltages, may be designated by $\{v_1, v_2, \ldots, v_N\}$, where $v_1 < v_2 < \ldots < v_N$. The set of corresponding logic values is denoted by $P = \{0,1,\ldots,p\}$, where $p = N-1$. The logic value $i \in P$ corresponds to the physical value v_{i+1}. The operation of addition (+) and multiplication (·) on P are defined by

$$\text{OR} \equiv \max(x,y) = x + y$$

$$\text{AND} \equiv \min(x,y) = x \cdot y, \quad x \text{ and } y \in P.$$

For any $x,y,z \in P$, and the constants 0 and p, the following properties hold:

IDEMPOTENCE:	$x + x = x$
	$x \cdot x = x$
COMMUTATIVITY:	$x + y = y + x$
	$x \cdot y = y \cdot x$
ASSOCIATIVITY:	$x + (y + z) = (x + y) + z$
	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$

ABSORPTION:	$x + (x \cdot y) = x$	
	$x \cdot (x + y) = x$	
DISTRIBUTIVITY:	$x + (y \cdot z) = (x + y) \cdot (x + z)$	
	$x \cdot (y + z) = (x \cdot y) + (x \cdot z)$	
NULL ELEMENTS:	$x + 0 = x$	$x \cdot p = x$
UNIVERSAL ELEMENTS:	$x + p = p$	$x \cdot 0 = 0$.

In an N-valued system, there are N^N unary (one variable) functions from P to P. Of special interest are the N functions defined by

$$X^i = \begin{cases} 0 \text{ if } x \neq i \\ p \text{ if } x = i, \; i \in P \end{cases}$$

and satisfying the relationship

$$X^0 + X^1 + \cdots + X^p = p.$$

Using the max and min operations and the N unary functions defined previously, any N-ary function of n variables, $f(\vec{x}) \equiv f(x_1,\ldots,x_n)$, can be represented in a canonical form

$$f(\vec{x}) = f_0(\vec{x}) + f_1(\vec{x}) + \cdots + f_p(\vec{x})$$

in which the function f_0 attains only the value 0, f_1 may be either 0 or 1,..., and f_p is either 0 or p. Table 1 represents a ternary (N = 3) 2-variable function. Using the canonical form for ternary functions, it may be written

$$\begin{aligned} f(x_1,x_2) &= X_1^0 \cdot X_2^1 + X_1^0 \cdot X_2^2 + X_1^1 \cdot X_2^2 + X_1^2 \cdot X_2^2 + \\ &\quad 1 \cdot (X_1^1 \cdot X_2^0 + X_1^2 \cdot X_2^0 + X_1^2 \cdot X_2^1) \\ &= X_1^0 \cdot X_2^1 + (X_1^0 + X_1^1 + X_1^2) \cdot X_2^2 + \\ &\quad 1 \cdot (X_1^1 \cdot X_2^0 + X_1^2 \cdot X_2^0 + X_1^2 \cdot X_2^1) \\ &= X_1^0 \cdot X_2^1 + X_2^2 + 1 \cdot (X_1^1 \cdot X_2^0 + X_1^2 \cdot X_2^0 + X_1^2 \cdot X_2^1). \end{aligned}$$

ADDERS

For a binary adder, we can assume that each input variable is available in both its true and complemented form. For example, the variable may be stored in a flip-flop whose output lines contain the variable and its complement. Note that the true and complemented values correspond to the two unary functions X^0 and X^1, respectively, for N = 2 and P = {0,1}. The output lines of the flip-flop can then be used as inputs to a combinational network of AND and OR gates which realizes an adder stage.

In base N we assume that the N unary functions of each variable described above are available as inputs to a combinational network of max and min gates which realizes the base N adder stage. We can think of these inputs as coming from a base N storage device, see Figure 1. If

$x = x_1x_2\ldots x_n$ and $y = y_1y_2\ldots y_n$ are two base N numbers, then stage i of the adder has the structure indicated in Figure 2.

The question of synchronous or asynchronous operation must be considered. In higher-order logics, this question is complicated by the necessity of deciding what ordering is to be imposed on the logic values. For an N-valued logic, three orderings are possible. Consider the set $P = \{0,1,\ldots,p\}$ of logic values with $0 < 1 < \cdots < p$, and any $i \in P$. In linear ordering, i is adjacent to the values i-1 and i+1 for $i \neq 0$ or p, and 0 is adjacent only to 1, and p only to p-1. In rotational ordering, i is adjacent to i-1 and to i+1 for all i; in particular, 0 is adjacent to p as well as 1, and p is adjacent to 0 and to p-1. In complete ordering, any $i \in P$ is adjacent to every other element in the set P. For N = 2, all three orderings are equivalent; for N = 3, the rotational and complete orderings are identical.

Devices that can shift from one voltage level, for example, to any one of a number of other levels must pass through one or more intermediate levels. Such a device satisfies the requirements of complete ordering but is unsuitable for asynchronous operation. Phase shift devices can be used for rotational ordering; but since they must pass through intermediate levels when changing from one level to another, when there are more than three levels and excluding the shift from the highest voltage to the lowest, they cannot be operated in an asynchronous mode. Linear devices have the property of linear ordering and are strictly synchronous in operation. These considerations combined with an evaluation of hazards in N-valued asynchronous circuits (for an analysis of hazards in ternary networks see [7] and [10]) lead us to consider only clocked adders.

With these preliminaries in mind, we can proceed with the cost analysis of the base N adders.

We assume that the storage devices used in the design of the base N adders are available at a cost that increases linearly as the radix increases. For example, if the base 2 device has a cost of 1 unit, the base 3 device costs 1.5 units, base 4 costs 2 units, etc. This is in accord with the classical argument of Burks, Goldstine, and von Neumann for the use of base 2 in the design of electronic computers [11]. It should be noted that their analysis actually shows that base e is the most economical base for the representation and storage of numbers but that base 2 is used because it is the most practical base nearest to e. This analysis does not consider the cost of logic circuitry. Thus although we will consider storage cost, our primary concern is the cost of the combinational logic needed to realize an adder stage. Since the combinational circuitry of the adder stage is composed entirely of max and min gates, logic cost will be a measure of the number of gates and inputs needed.

As a basis for comparison we will use the hypothetical logic cost ratio $\log_{10}N/\log_{10}2$, see Table 2. This ratio is our break-even logic cost bound, that is, if the actual logic cost for a particular base is less than this ratio, then an adder in this base has a lower logic cost than the base 2 adder for the same range of operands considered.

Let M be the magnitude of the largest number the adder must accommodate. Then we have

$$2^x = M \qquad N^y = M$$

$$x \log_{10} 2 = \log_{10} M \qquad y \log_{10} N = \log_{10} M$$

$$x = \frac{\log_{10} M}{\log_{10} 2} \qquad y = \frac{\log_{10} M}{\log_{10} N}$$

where x is the number of binary adder stages needed and y is the number of N-ary adder stages needed.

Assuming that storage cost is linearly increasing with the base, we have

Base 2 Storage Cost: $SC_2 = 1x$

Base N Storage Cost: $SC_N = \left(\frac{N}{2}\right)y$

$$\frac{SC_N}{SC_2} = \left(\frac{N}{2}\right)\frac{\log_{10} 2}{\log_{10} N} .$$

This ratio gives the relative cost of base N storage compared to base 2 storage, see Table 3. The values of x and y are also used to calculate the binary and N-ary logic costs as

Binary Cost: $C_2 = x(aG_2 + bI_2) = \left(\frac{\log_{10} M}{\log_{10} 2}\right)(aG_2 + bI_2)$

N-ary Cost: $C_N = y(aG_N + bI_N) = \left(\frac{\log_{10} M}{\log_{10} N}\right)(aG_N + bI_N)$

where G_2 = number of gates per binary adder stage

G_N = number of gates per N-ary adder stage

I_2 = number of inputs per binary adder stage

I_N = number of inputs per N-ary adder stage

a = gate cost factor

b = input cost factor.

Thus the logic cost ratio used to compare with the break-even logic cost bound is

$$\frac{C_N}{C_2} = \left(\frac{aG_N + bI_N}{\log_{10} N}\right)\left(\frac{\log_{10} 2}{aG_2 + bI_2}\right) .$$

As an aid in determining the number of gates and inputs needed in a base N adder stage, we consider the development of general formulas for the sum, S_i, and carry, C_{i+1}, functions.

Since the value of a carry in any base is either 0 or 1, we examine the cases $C_i = 0$ and $C_i = 1$, where C_i is the carry to stage i of the adder from stage i-1. Let x_i and y_i denote the i^{th} digits of the addend and the augend, respectively.

Case 1: $C_i = 0$.

S_i can take on any value between 0 and N-1.

Then $S_i = N-1$ for $x_i = 0$ and $y_i = N-1$;

...; $x_i = N-1$ and $y_i = 0$,

$S_i = N-2$ for $x_i = 0$ and $y_i = N-2$;

...; $x_i = N-1$ and $y_i = N-1$; etc.

Using the N unary functions we can write

$$S_i = C_i^0 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-1)-j} + (N-2) \cdot C_i^0 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-2)-j} + \cdots$$
$$\cdots + 1 \cdot C_1^0 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{1-j}$$

where 1) all arithmetic in the superscripts is modulo N,

2) $-0 \equiv 0$.

Case 2: $C_i = 1$.

S_i can take on any value between 0 and N-1.

Then $S_i = N-1$ for $x_i = 0$ and $y_i = N-2$;

...; $x_i = N-1$ and $y_i = N-1$,

$S_i = N-2$ for $x_i = 0$ and $y_i = N-3$;

...; $x_i = N-1$ and $y_i = N-2$; etc.

We write this as

$$S_i = C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-1)-1-j} + (N-2) \cdot C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-2)-1-j} + \cdots$$
$$\cdots + 1 \cdot C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{-j}$$

with the same rules as in Case 1.

Combining the results of Case 1 and Case 2 and noting that

$C_i = 0$: $S_i = N-2, N-3, \ldots, 1$

$C_i = 1$: $S_i = N-1, N-2, \ldots, 2$,

we can write the generalized sum function as

$$S_i = C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-1)-1-j} + C_i^0 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-1)-j} +$$
$$(N-2) \cdot \left[C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{(N-2)-1-j} + \sum_{j=0}^{N-1} X^j \cdot Y^{(N-2)-j} \right] +$$
$$\vdots$$
$$+ 1 \cdot \left[C_i^1 \cdot \sum_{j=0}^{N-1} X^j \cdot Y^{-j} + \sum_{j=0}^{N-1} X^j \cdot Y^{1-j} \right].$$

Similar analysis yields the generalized carry function

$$C_{i+1} = 1\cdot[C_i^1\cdot\sum_{j=0}^{N-1} X^j\cdot Y^{(N-1)-j} + \sum_{j=1}^{N-1} X^j\cdot Y^{(N)-j}$$
$$+ \sum_{j=2}^{N-1} X^j\cdot Y^{(N+1)-j} + \cdots + \sum_{j=N-1}^{N-1} X^j\cdot Y^{(2N-2)-j}]$$

where the constant 1 is dropped for N = 2.

An example, for N = 3 the formulas become

$$S_i = C_i^1\cdot\sum_{j=0}^{2} X^j\cdot Y^{(2)-1-j} + C_i^0\cdot\sum_{j=0}^{2} X^j\cdot Y^{(2)-j} +$$
$$1\cdot[C_i^1\cdot\sum_{j=0}^{2} X^j\cdot Y^{(1)-1-j} + \sum_{j=0}^{2} X^j\cdot Y^{(1)-j}]$$
$$= C_i^1\cdot(X^0\cdot Y^1 + X^1\cdot Y^0 + X^2\cdot Y^2) + C_i^0\cdot(X^0\cdot Y^2 + X^1\cdot Y^1 + X^2\cdot Y^0) +$$
$$1\cdot[C_i^1\cdot(X^0\cdot Y^0 + X^1\cdot Y^2 + X^2\cdot Y^1) + X^0\cdot Y^1 + X^1\cdot Y^0 + X^2\cdot Y^2]$$

$$C_{i+1} = 1\cdot[C_i^1\cdot\sum_{j=0}^{2} X^j\cdot Y^{(2)-j} + \sum_{j=1}^{2} X^j\cdot Y^{(3)-j} + \sum_{j=2}^{2} X^j\cdot Y^{(4)-j}]$$
$$= 1\cdot[C_i^1\cdot(X^0\cdot Y^2 + X^1\cdot Y^1 + X^2\cdot Y^0) + X^1\cdot Y^2 + X^2\cdot Y^1 + X^2\cdot Y^2].$$

Using these formulas the number of gates and inputs per stage can be computed. For base 2, the formulas give

$$S_i = C^1\cdot X^0\cdot Y^0 + C^1\cdot X^1\cdot Y^1 + C^0\cdot X^0\cdot Y^1 + C^0\cdot X^1\cdot Y^0$$
$$C_{i+1} = C^1\cdot X^0\cdot Y^1 + C^1\cdot X^1\cdot Y^0 + X^1\cdot Y^1 .$$

These can be factored in two ways:

1) $C_{i+1} = C^1\cdot\underbrace{(X^0\cdot Y^1 + X^1\cdot Y^0)}_{B} + \underbrace{X^1\cdot Y^1}_{A}$

$$S_i = C^1\cdot X^0\cdot Y^0 + C^1\cdot A + C^0\cdot B$$

2) $C_{i+1} = C^1\cdot(Y^1 + X^1) + \underbrace{X^1\cdot Y^1}_{A}$

$$S_i = C^1\cdot X^0\cdot Y^0 + C^1\cdot A + C^0\cdot X^0\cdot Y^1 + C^0\cdot X^1\cdot Y^0.$$

The first realization has a cost of 10 gates and 22 inputs while the second costs 9 gates and 23 inputs.

The functions for base 3 can be written

$$C_{i+1} = 1\cdot[C^1\cdot\underbrace{(X^0\cdot Y^2 + X^1\cdot Y^1 + X^2\cdot Y^0)}_{A} + \underbrace{X^1\cdot Y^2}_{E} + \underbrace{X^2\cdot Y^1}_{D} + \underbrace{X^2\cdot Y^2}_{B}] .$$

$$S_i = C^1 \cdot \underbrace{(X^0 \cdot Y^1 + X^1 \cdot Y^0 + B)}_{F} + C^0 \cdot A + 1 \cdot [C^1 \cdot (X^0 \cdot Y^0 + D + E) + F].$$

Implementation of these expressions has a cost of 21 gates and 48 inputs. Similarly, the costs for the other bases are arrived at. Table 4 summarizes the results for N = 2 through 8. It should be noted that the technique of term sharing that is used in simplifying the sum and carry functions in all bases does not yield a two-level realization of a base N adder. Analysis shows that two-level realizations result in adders that are, in general, much more costly than the base 2 adder. Table 5 indicates the values of the logic cost ratio C_N/C_2 and the normalized ratio $C_N/C_2/{}^{\log_{10}N}/_{\log_{10}2}$ for several values of the cost factors a and b. The underlined entries in Table 5 are those ratios which are less than or equal to the break-even logic cost bound. It is evident that bases 3, 4, and 5 have a lower logic cost than base 2 and that base 6, in some cases, is also less costly. Further, even after taking into account the relative storage costs, Table 3, we see that bases 3 and 4 have less costly implementations of the adder than base 2 for all values of the cost factors considered. Base 5 is no more expensive than base 2 for some values of the cost factors, but base 6 becomes more expensive in all cases when storage cost is considered.

CONCLUSION

We have presented an application of N-valued combinational switching theory which shows that the use of higher bases can be very advantageous. The argument that base 2 is always the most economical base to use needs to be reexamined in the light of developments in multi-valued switching theory. We feel that further investigation of the theory and uses of both combinational and sequential N-valued switching theory is warranted.

REFERENCES

1. Berlin, R. D., "Synthesis of N-valued Switching Circuits," IRE Transactions on Electronic Computers, Vol. EC-7, pp. 52-66, March, 1958.

2. Lee, C. Y. and W. H. Chen, "Several Valued Combinational Switching Circuits," Transactions of the AIEE (Communication and Electronics), Vol. 75, pp. 278-283, July, 1956.

3. Lowenschuss, O., "Nonbinary Switching Theory," IRE National Convention Record, Vol. 1, pt. 4, pp. 305-317, March, 1958.

4. Givone, D. D. and R. W. Snelsire, "The Design of Multiple-Valued Logic Systems," State University of New York at Buffalo, Digital Systems Laboratory of the Department of Electrical Engineering, June, 1968.

5. Yoeli, M. and G. Rosenfeld, "Logical Design of Ternary Switching Circuits," IEEE Transactions on Electronic Computers, Vol. EC-14, pp. 19-29, February, 1965.

6. Santos, J. and M. Arango, "On the Analysis and Synthesis of Three-Valued Digital Systems," AFIPS Conference Proceedings, Vol. 25, Spring Joint Computer Conference, 1964.

7. Wojcik, A. S. and G. Metze, "A Reexamination of Ternary Switching Theory," Proceedings of the Sixth Annual Allerton Conference on Circuit and System Theory, pp. 711-721, October, 1968.

8. Post, E. L., "Introduction to a General Theory of Elementary Propositions," American Journal of Mathematics, Vol. 43, pp. 163-185, 1921.

9. Rosenbloom, P. C., The Elements of Mathematical Logic, Dover, New York, 1950.

10. Abecassis, L., "A Study of Hazards in Ternary Switching Circuits," University of Illinois, Coordinated Science Laboratory Report, R-310, August, 1966.

11. Burks, A. W., H. H. Goldstine, and J. von Neumann, "Preliminary Discussion of the Logical Design of an Electronic Computing Instrument," Institute for Advanced Study, Princeton, N. J., 1st Ed. June, 1946, 2nd Ed., 1947, Sec. 5.2.

x_1	x_2	f
0	0	0
0	1	2
0	2	2
1	0	1
1	1	0
1	2	2
2	0	1
2	1	1
2	2	2

Table 1. Two-Variable Ternary Function

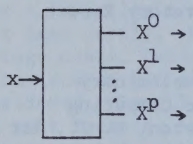
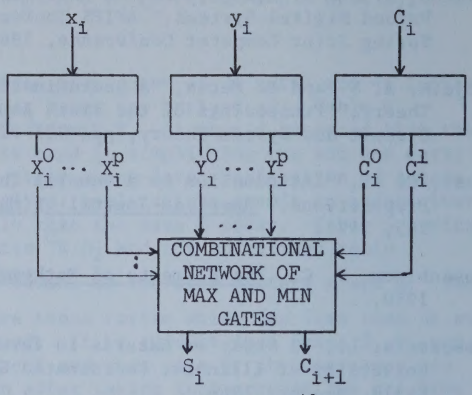


Figure 1. Base N Storage Device



x_i, y_i, S_i denote the i^{th} base N digit of the addend, augend, and sum, respectively

C_i denotes the carry signal into stage i

C_{i+1} denotes the carry signal from stage i to stage i+1

Figure 2. Base N Adder Stage

N	$\frac{\log_{10} N}{\log_{10} 2}$
3	1.58
4	2.00
5	2.32
6	2.58
7	2.81
8	3.00

Table 2. Hypothetical Cost Ratio: $\frac{\log_{10} N}{\log_{10} 2}$

N	$\frac{SC_N}{SC_2}$
2	1.00
3	.95
4	1.00
5	1.08
6	1.16
7	1.25
8	1.33

Table 3. Relative Cost of Base N Storage to Base 2 Storage: SC_N/SC_2

N	G_N	I_N
2	9	23
	10	22
3	21	48
4	33	78
5	47	114
6	63	156
7	81	204
8	101	258

Table 4. Gates and Inputs Per Base N Adder Stage

a	b	N	C_N/C_2	$\frac{C_N/C_2}{\log_{10}N/\log_{10}2}$
1	0	3	1.47	.930
		4	1.83	.915
		5	2.25	.970
		6	2.71	1.050
		7	3.21	1.142
		8	3.74	1.247
2	1	3	1.39	.880
		4	1.76	.880
		5	2.18	.940
		6	2.66	1.031
		7	3.18	1.132
		8	3.74	1.247
10	1	3	1.44	.911
		4	1.81	.905
		5	2.23	.961
		6	2.69	1.043
		7	3.20	1.139
		8	3.74	1.247
0	1	3	1.32	.835
		4	1.70	.850
		5	2.13	.918
		6	2.62	1.016
		7	3.16	1.125
		8	3.74	1.247
1	10	3	1.32	.835
		4	1.71	.855
		5	2.14	.922
		6	2.63	1.019
		7	3.16	1.125
		8	3.74	1.247
1	1	3	1.36	.861
		4	1.73	.865
		5	2.16	.931
		6	2.65	1.027
		7	3.17	1.128
		8	3.74	1.247

a) $G_2 = 9$, $I_2 = 23$

a	b	N	C_N/C_2	$\frac{C_N/C_2}{\log_{10}N/\log_{10}2}$
1	0	3	1.33	.842
		4	1.65	.825
		5	2.02	.871
		6	2.44	.946
		7	2.89	1.028
		8	3.37	1.123
2	1	3	1.35	.854
		4	1.71	.855
		5	2.13	.918
		6	2.60	1.008
		7	3.10	1.103
		8	3.65	1.217
10	1	3	1.33	.842
		4	1.67	.835
		5	2.06	.888
		6	2.49	.965
		7	2.96	1.053
		8	3.46	1.153
0	1	3	1.38	.873
		4	1.77	.885
		5	2.23	.961
		6	2.74	1.062
		7	3.30	1.174
		8	3.91	1.303
1	10	3	1.37	.867
		4	1.77	.885
		5	2.22	.957
		6	2.73	1.058
		7	3.28	1.167
		8	3.89	1.297
1	1	3	1.36	.861
		4	1.73	.865
		5	2.16	.931
		6	2.65	1.027
		7	3.17	1.128
		8	3.74	1.247

b) $G_2 = 10$, $I_2 = 22$

Table 5. Values of C_N/C_2 and $C_N/C_2 \Big/ \log_{10}N/\log_{10}2$

A MARKOV MODEL FOR RELIABILITY OF PROBABILISTIC SEQUENTIAL CIRCUITS WITH MARKOV INPUTS*.

G.S. GLINSKI and WHO-KEE CHUNG,
Department of Electrical Engineering,
University of Ottawa,
Ottawa, Canada.

1. ABSTRACT

This paper is concerned with an extension of a Markov model of probabilistic sequential circuit (p.s.c.) to the case when input is also Markovian. A simpler method of obtaining the total transition matrix of Markov model of p.s.c.'s with q-th order ($q \geq 1$) Markov chain input is developed and its reliability is estimated.

2. INTRODUCTION

The use of probabilistic logical matrix (P-matrix, [P]) and the idea of total transition matrix, $[P_T]$, to estimate the reliability of a p.s.c. using a Markov chain model with first order Markov chain input was first developed by Udagawa and Inagaki [1]. Also a simpler Markov model was formed and its reliability was calculated, utilizing z-transform and signal flow-graph techniques, by Glinski and Chung [2] for the p.s.c. with Bernoulli inputs. The problem considered in this paper is to obtain the Markov model and calculate the reliability of a p.s.c. with Markov inputs. Here the reliability of a p.s.c. is considered in terms of the mean-time-to-first-error, T_m.

3. FUNDAMENTALS

3.1. P-Matrix, Series, Parallel and "Side by Side" Connections of N Circuits [2].

3.2. Enlargement of P-Matrix (Mixed Combination of Circuits).

The circuits 1 and 2 of Fig. 1 are partially "in parallel" and "side by side", thus none of the operations introduced in ref. 2 can be used directly. However the parallel operation can be applied if the P-matrices of both circuits 1 and 2, denoted by $[P_1]$ and $[P_2]$, are properly modified. The modified P-matrices will have 2^p rows and the number of columns is unchanged. The enlarged P-matrices for circuits 1 and 2 are denoted by $[P_1]_e$ and $[P_2]_e$ respectively. The entries in $[P_1]_e$ are the same as the entries in $[P_1]$ when the valuations of $(x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_n)$'s are the same for both $[P_1]$ and $[P_1]_e$. Similarly we can obtain $[P_2]_e$. The P-matrix of the combined circuit is

$$[P] = [P_1]_e * [P_2]_e$$

The "*" denotes the parallel operation.

4. MARKOV MODEL FOR PROBABILISTIC SEQUENTIAL CIRCUITS WITH MARKOV INPUTS.

In ref. 2, we introduce a new column and a new row, denoted by e(k+1) and e(k) respectively, in the $[P_s]$, (P-matrix of p.s.c.). We can get the modified P-matrix of the p.s.c., $[P_{ce}]$, where the entries in column e(k+1) contains all the probabilities of erroneous operations of the p.s.c. The probabilities of correct operations of the p.s.c. remain unchanged.

* This work was supported in part by the National Research Council of Canada, Grant No. A-875.

Now consider the case when transitions of inputs to the p.s.c. are represesented by a first order Markov chain, denoted by $[P_{I1}]$. From $[P_{I1}]$ we know that the probability of next input state $j, j \in \{0, 1, \ldots, 2^n - 1\}$ where n is the number of input variables to the p.s.c., depends only on the present input state. We can modify $[P_{I1}]$ such that the row headings of $[P_{I1}]$ are the same as $[P_{ce}]$, the modified $[P_{I1}]$ is denoted by $[P_{I1}]_e$. The parallel operation can now be applied to these P-matrices, $[P_{I1}]_e$ and $[P_{ce}]$, thus the total transition matrix is

$$[P_{TM}] = [P_{I1}]_e * [P_{ce}]$$

Similarly we can get the total transition matrix when transitions of inputs are represented by a q-th order ($q > 1$) Markov chain, denoted by $[P_{Iq}]$. The total transition matrix is

$$[P_{TM}] = [P_{Iq}]_e * [P_{ce}]_e$$

5. PROCEDURE FOR FINDING THE RELIABILITY OF P.S.C. WITH MARKOV INPUTS.

The procedure for finding the reliability of p.s.c.'s with Markov inputs is shown below:

1. Find $[P_s]$, then find $[P_{ce}]$ from $[P_s]$.
2. Enlarge $[P_{I1}]$ or $[P_{Iq}]$. If transitions of inputs are represented by a second or higher order Markov chain, then enlarge $[P_{ce}]$. Find $[P_{TM}]$.
3. Draw the z-transform of the state diagram of correct transitions. Calculate T_m's.

6. CONCLUSION

A simplified and systematic way for finding the Markov model of p.s.c.'s with Markov inputs is presented. The dimension of it is smaller than the one previously developed [1]. This model is more general than those [1], [2] developed before.

7. REFERENCES

1. K. Udagawa and Y. Inagaki, "Probabilistic studies of sequential circuits and estimation of their reliability using Markov chain", J. of the Inst. of Elect. Comm. Engg. in Japan, vol. 47, No. 4, pp. 276-286, April 1964.
2. G.S. Glinski and Who-Kee Chung, "A Markov model for reliability of probabilistic sequential circuit with Bernoulli inputs", Proc. of the Sixth Annual Allerton Conf. on Circuit and System Theory, pp. 568-577, Oct. 2-4, 1968.

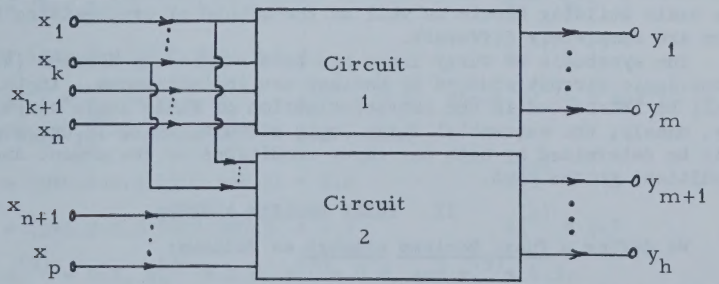


Fig. 1. Two switching circuits with some common input variables.

ANALYSIS AND SYNTHESIS OF SEQUENTIAL FUZZY LOGIC CIRCUITS*

SAMUEL C. LEE
Department of Electrical Engineering
New York University
Bronx, New York 10453

SUMMARY

Fuzzy logic deals with propositions which may be ascribed subjectively values between falsehood and truth (zero and one) either in a continuous or a discrete fashion. This is in contrast with ordinary logic (two-valued or many-valued logic) in which a given proposition is ascribed values objectively using either deterministic or probabilistic approaches. A sequential fuzzy logic circuit is a dynamic fuzzy logic circuit whose outputs will be determined by both the input conditions of the moment and the input conditions in the past. A detailed study of the analysis and synthesis of sequential fuzzy logic circuits is presented and illustrated by examples.

I. INTRODUCTION

The concept of fuzziness was introduced by Zadeh [1]-[3]. Fuzzy logic [4] deals with propositions which may be ascribed subjectively values between falsehood and truth (zero and one) either in a continuous or a discrete fashion. This is in contrast with ordinary logic (two-valued or many-valued logic) in which a given proposition is ascribed values objectively using either deterministic or probabilistic approaches. A nonfuzzy statement is as such

"If John passes both examination A and examination B, then the company which he works with will pay his tuition if and only if the company has money available and his boss approves it."

An appropriate logical analysis of this statement might begin with the symbolism

$$a \wedge b \supset (p \equiv m \wedge f) \tag{1}$$

where the letters denote the following statements

a: John passes examination A, m: The company has money available
b: John passes examination B, f: His boss approves it.
p: The company which he works with will pay his tuition,

The proposition in Eq. (1) can be synthesized by an ordinary logic circuit. If, however, the above statement is changed to

"If John passes both examination A and examination B satisfactorily, the company which he works with will pay a large portion of his tuition if and only if the company has a large amount of uncommitted money and his boss approves it without too much reservation."

which becomes a fuzzy statement. Even though the logical expression of this statement is identical to that of the previous statement, namely, Eq. (1), the basic building blocks as well as the method of synthesizing this expression are completely different.

The synthesis of fuzzy logic has been studied by Marinos [4]. The fuzzy logic circuit studied by Marinos are instantaneous. In this paper, we shall be interested in the characterization of fuzzy logic circuit with memory, namely, the sequential fuzzy logic circuit, which is one whose outputs will be determined by both the input conditions of the moment and the input conditions in the past.

II. FUZZY BOOLEAN ALGEBRA

We define a fuzzy Boolean algebra as follows:

*The work is supported in part by the National Science Foundation Grant GK-3791

Definition. A fuzzy Boolean algebra is an algebraic system $(B,+,\cdot,\sim)$ which consists of (1) A buzzy set B,(the grade of membership of every member of B is between 0 and 1); (2) Two binary operations + and · which are defined as: If x,y in B, $x+y=\max(x,y)$ and $x\cdot y=\min(x,y)$; and (3) A unary operation $\sim$ defined as: If x in B, $\sim x=1-x$. For convenience, we shall write $\sim x$ as x'.

The fuzzy set is closed under the three operations +,·, and $\sim$, which are associative and commutative and each distributes over the other.

From this definition, the following consequences are immediate [4]. Let B be a fuzzy Boolean algebra, x,y,z be elements in B. Then

$$x + x = x \qquad (x+y)' = x'z'$$
$$x \cdot x = x \qquad (xz)' = x' + z'$$
$$(xy)z = x(yz) \qquad x + x' = \max(x,x')$$
$$(x+z) + z = x + (y+z) \qquad x \cdot x' = \min(x,x')$$
$$x(y+z) = xy + xz \qquad x(x+y) = x$$

Definition -Fuzzy Logic Function. Let B be a fuzzy Boolean algebra. A fuzzy logic function defined on B is a mapping from B to B using exclusively the three operations +,·, and $\sim$.

For example, $f(x,y,z) = xy'zz' + xy'z' + x'yz'$ is a fuzzy logic function which may be simplified as

$$f(x,y,z) = xy'z'(z+1) + x'yz'$$
$$= xy'z' + x'yz'$$
$$= z'(xy'+x'y)$$

It should be noted that terms such as $xyzz'$ which reduces to zero in ordinary two-valued logic are not necessarily zero in fuzzy logic. Simiarly, factors such as $(x+x')z$ cannot, in general, be reduced to z.

III. ANALYSIS OF SEQUENTIAL FUZZY LOGIC CIRCUITS

Unlike the ordinary sequential logic circuit of which a relay contact is either on or off and a relay coil is either excited or not excited, in a sequential fuzzy logic circuit both relay contacts and relay coils may have any value between 0 and 1, so may have the input values. The analysis of a sequential fuzzy logic circuit is best illustrated by use of an example.

Consider the sequential fuzzy logic (SFLC) circuit of Fig. 1. The transmissions to relays Y_1 and Y_2 are

$$Y_1 = (y_1+x_2'y_2z_1')(x_1+y_2) \qquad Y_2 = (z_1'+x_2'y_2z_1)(x_1+z_2)$$

and the outputs are $Z_1 = y_2$ and $Z_2 = z_2'$.

As mentioned above, the input X_1, X_2 may be of any value between 0 and 1. Suppose $X_1=0.3$ and $X_2=0.8$. Then

Step 1 $x_1^{(1)} = 0.3$ and $x_2^{(1)} = 0.8$

$$Y_1^{(1)} = (0+0.2\cdot0.1)(0.3+0) = 0 \qquad Z_1^{(1)} = 0$$
$$Y_2^{(1)} = (1+0.2\cdot0.0)(0.3+0) = 0.3 \qquad Z_2^{(1)} = 1$$

Step 2 $x_1^{(2)} = x_1^{(1)} = 0.3$, $x_2^{(2)} = x_2^{(1)} = 0.8$. Since $Y_1^{(1)} = 0$ and $Y_2^{(1)} = 0.3$, they imply that $y_1^{(2)} = 0$ and $y_2^{(2)} = 0.3$. Thus

$$Y_1^{(2)} = (0+0.2\cdot0.3\cdot1)(0.3+0.3) = 0.2 \qquad Z_1^{(2)} = 0.3$$
$$Y_2^{(2)} = (1+0.2\cdot0.3\cdot0)(0.3+0.3) = 0.3 \qquad Z_2^{(2)} = 0.7$$

Step 3 $x_1^{(3)} = 0.3$, $x_2^{(3)} = 0.8$, $y_1^{(3)} = 0.2$, and $y_2^{(3)} = 0.3$.

$$Y_1^{(3)} = (0.2+0.2\cdot0.7\cdot0.8)(0.3+0.3) = 0.2 \qquad Z_1^{(3)} = 0.3$$

$$Y_2^{(3)} = (0.8+0.2\cdot0.3\cdot0.2)(0.3+0.3) = 0.3 \qquad Z_2^{(3)} = 0.7$$

Step 4 $x_1^{(4)} = 0.3$, $x_2^{(4)} = 0.8$, $y_1^{(4)} = 0.2$, and $y_2^{(4)} = 0.3$

$$Y_1^{(4)} = 0.2 \qquad Z_1 = 0.3$$

$$Y_2^{(4)} = 0.3 \qquad Z_2 = 0.7$$

The significance of this analysis can better be appreciated by sumarizing the steps we have described, as shown in the table below.

Step Number	Relay Coil		Relay		Contact		Relay Coil		Outputs	
	X_1	X_2	x_1	x_2	y_1	y_2	Y_1	Y_2	Z_1	Z_2
1	0.3	0.8	0.3	0.8	0	0	0	0.3	0	1
2	0.3	0.8	0.3	0.8	0	0.3	0.2	0.3	0.3	0.7
3	0.3	0.8	0.3	0.8	0.2	0.3	0.2	0.3	0.3	0.7
4	0.3	0.8	0.3	0.8	0.2	0.3	0.2	0.3	0.3	0.7

It is seen that there is no change from Step 3 to Step 4. Of course, the circuit will remain unchanged from Step 3 on. We shall call Step 3 a steady state of the circuit.

Definition A sequential fuzzy logic circuit is called a stable sequential fuzzy logic circuit (SSFLC) if for any inputs $0 \le X_i \le 1$, the circuit has a steady state, otherwise it is an unstable sequential fuzzy logic circuit.

From now on, our interest will be confined to SSFLC. It is clear that, in general, a SFLC cannot be ensured to have a steady state for all possible inputs. However, it can be shown that the class of SFLC whose general schematic diagram is shown in Fig. 2 are always stable.

Proof Suppose the inputs are $X_i = a_i$, $0 \le a_i \le 1$. It follows that $x_i = a_i$. From the diagram, it is seen that they are fixed once the inputs are fixed. The transmissions to relays Y_i and Z_i may be generally expressed as

$$Y_i = f_i(x,y) \qquad Z_i = g_i(x,y)$$

Now let us consider Y_i. It is easy to show that they may be expressed in terms of a product of a pxq matrix and a qxl matrix, denoted by x^*_Y and y^*_Y, respectively. The matrix y^* has q elements, some of the elements of the set of $y_1, y_2, \ldots, y_p, y_1', y_2', \ldots, y_p'$, plus possible combinations of y_i and y_i'. The elements of x^* are the corresponding multiplicative coefficients of the elements of y^*. Using the above example,

$$Y_1 = (y_1 + x_2' y_2 y')(x_1 + y_2) = x_1 y_1 + x_2' y_1' y_2 + y_1 y_2$$

$$Y_2 = (y_1' + x_2' y_2 y_1)(x_1 + y_2) = x_1 y_1' + y_1' y_2 + x_2' y_1 y_2$$

therefore we may write

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \quad \begin{bmatrix} x_1 & 0 & x_2' & 1 \\ 0 & x_1 & 1 & x_2' \end{bmatrix} \begin{bmatrix} y_1 \\ y_1' \\ y_1'y_2 \\ y_1y_2 \end{bmatrix}$$

where

$$x_Y^* = \begin{bmatrix} x_1 & 0 & x_2' & 1 \\ 0 & x_1 & 1 & x_2' \end{bmatrix} \quad \text{and} \quad y_Y^* = \begin{bmatrix} y_1 \\ y_1 \\ y_1'y_2 \\ y_1y_2 \end{bmatrix}$$

Similarly,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_2 \\ y_2' \end{bmatrix}$$

where

$$x_Z^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad y_Z^* = \begin{bmatrix} y_2 \\ y_2' \end{bmatrix}$$

Since at Step 2, every element of y* are determined by $x_1^{(1)}, x_2^{(1)}, \ldots, x_m^{(1)}$ which, as mentioned before, for this class of circuits, are determined and fixed by the inputs to the circuit, i.e., $x_i^{(n)}=x_i^{(1)}=X_i$. Thus $Y_i^{(2)}$, i=1,2,...,p, are functions of $x_i^{(1)}$, i=1,2,...,m only which are fixed values. Hence $Y_i^{(n)}=Y_i^{(2)}$, for $n\geq 2$. Again, use the above example

$$\begin{bmatrix} Y_1^{(1)} \\ Y_2^{(1)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & 0 & (x_2^{(1)})' & 1 \\ 0 & x_1^{(1)} & 1 & (x_2^{(1)})' \end{bmatrix} \begin{bmatrix} y_1^{(1)} \\ (y_1^{(1)})' \\ (y_1^{(1)})'y_2^{(1)} \\ y_1^{(1)}y_2^{(1)} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^{(1)} & 0 & (x_2^{(1)})' & 1 \\ 0 & x_1^{(1)} & 1 & (x^{(1)})' \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1\cdot 0 \\ 0\cdot 0 \end{bmatrix} = \begin{bmatrix} 0 \\ x^{(1)} \end{bmatrix}$$

$$\begin{bmatrix} Y_1^{(2)} \\ Y_2^{(2)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & 0 & (x_2^{(1)})' & 1 \\ 0 & x_1^{(1)} & 1 & (x_2^{(1)})' \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \cdot x_1^{(1)} \\ 0 \cdot x_1^{(1)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)}(x_2^{(1)})' \\ x_1^{(1)} \end{bmatrix}$$

When $x_1^{(1)}=0.3$ and $x_2^{(1)}=0.8$, $Y_1^{(2)}=0.2$ and $Y_2^{(2)}=0.3$, which agree with the results obtained previously.

IV. SYNTHESIS OF SEQUENTIAL FUZZY LOGIC CIRCUITS

From above, it is seen that a fuzzy function does not provide us with a "decision mechanism". For instance, when a fuzzy logic function $f(x_1,x_2)=a$, $0<a<1$, it does not have much significance unless it is related to some outcome which dependent upon the fact that $f(x_1,x_2)=a$. A way of introducing this quality into a fuzzy system was introduced by Marinos[4]. He suggested to reintroduce the concept of "belonging to a set" but in a rather qualified sense by subdividing the continuous range of membership grades [0,1] into a finite number of classes:

$$\begin{aligned} &\text{Class 1:} \quad \alpha_1 \leq x \leq 1 \\ &\text{Class 2:} \quad \alpha_2 \leq x < \alpha_1 \\ &\quad\vdots \qquad\qquad \vdots \\ &\text{Class n:} \quad 0 \leq x < \alpha_{n-1} \end{aligned}$$

where $0<\alpha_i<1$ and $\alpha_i>\alpha_j$, if $i<j$. The synthesis problem we shall be concerned with will be based on such a classification of the continuous range of membership grades.

The synthesis procedure is best illustrated by the use of an example.

A sequential fuzzy logic circuit of the form indicated in Fig. 3 is to be designed so that if $X_2>\alpha$, $0<\alpha<1$, the output Z is the same as the X_1 input but if $X_2\leq\alpha$, the output is to remain fixed at its last value prior to X_2 becoming $\leq \alpha$. No more than one input variable is to change at one time.

For the reason mentioned before, we subdivide the continuous range of membership grades of X_1 into n classes. The simplest case would be n=z and

$$\begin{aligned} &\text{Class 1:} \quad \alpha < X_1 \leq 1 \\ &\text{Class 2:} \quad 0 \leq X_1 \leq \alpha \end{aligned}$$

Define $X_i=\alpha^+$ if $\alpha < X_i \leq 1$, and $X_i = \alpha^-$ if $0 < X_i \leq \alpha$.

Then we can easily construct the primitive flow table from the design specification as shown in Table 2. The merged flow table is given in Table 3.

Table 2

x_1x_2 $\alpha^-\alpha^-$	$\alpha^-\alpha^+$	$\alpha^+\alpha^+$	$\alpha^+\alpha^-$	Z
(1)	3	-	5	0
(2)	3	-	6	1
1	(3)	4	-	0
-	3	(4)	6	1
1	-	4	(5)	0
2	-	4	(6)	1

Table 3

x_1x_2	$\alpha^-\alpha^-$	$\alpha^-\alpha^+$	$\alpha^+\alpha^+$	$\alpha^+\alpha^-$
	①	③	4	⑤
	②	3	④	⑥

Since there are only two rows in the merged flow table, only one secondary relay is needed. Denote it by Y. Suppose Y is subdivided in the same way as X_1 and X_2. In this case, it is very similar to the case where the circuit is an ordinary sequential circuit. From Table 3, one way of making the secondary assignment is shown in Table 4.

Table 4

y \ x_1x_2	$\alpha^-\alpha^-$	$\alpha^-\alpha^+$	$\alpha^+\alpha^+$	$\alpha^+\alpha^-$
α^-	α^-	α^-	α^+	α^-
α^+	α^+	α^-	α^+	α^+

Y

y \ x_1x_2	$\alpha^-\alpha^-$	$\alpha^-\alpha^+$	$\alpha^+\alpha^+$	$\alpha^+\alpha^-$
α^-	α^-	α^-	-	α^-
α^+	α^+	-	α^+	α^+

Z

In accordance to Table 4, the fuzzy logic functions for the secondary relay and output are

$$Y = x_1x_2y' + x_1x_2'y + x_1'x_2'y + x_1x_2'y'$$

$$Z = y$$

The realization[see Ref. 4] is shown in Fig. 4. In general, for 2-class fuzzy logic,

$$X_1: \begin{cases} \text{class 1: } \alpha < X_1 \le 1 \\ \text{class 2: } 0 \le X_1 \le \alpha \end{cases}$$

$$X_2: \begin{cases} \text{class 1: } \beta < X_2 \le 1 \\ \text{class 2: } 0 \le X_2 \le \beta \end{cases}$$

$$Y: \begin{cases} \text{class 1: } \gamma < Y \le 1 \\ \text{class 2: } 0 \le Y \le \gamma \end{cases}$$

the excitation and output matrices are shown in Table 5.

Table 5

y \ x_1x_2	$\alpha^-\beta^-$	$\alpha^-\beta^+$	$\alpha^+\beta^+$	$\alpha^+\beta^-$
γ^-	γ^-	γ^-	γ^+	γ^-
γ^+	γ^+	γ^-	γ^+	γ^+

Y

y \ x_1x_2	$\alpha^-\beta^-$	$\alpha^-\beta^+$	$\alpha^+\beta^+$	$\alpha^+\beta^-$
γ^-	γ^-	γ^-	—	γ^-
γ^+	γ^+	—	γ^+	γ^+

Z

This realization is shown in Fig. 5.

V. CONCLUSION

The analysis and synthesis of sequential fuzzy logic circuits has been investigated. The synthesis procedure described above can be extended to synthesizing SSFLC with

(1) multiple secondary relays, i.e., $Y_1, Y_2, \ldots, Y_p$,

(2) multiply outputs, i.e., $Z_1, Z_2, \ldots, Z_n$,

(3) all primary and secondary relays being n-class fuzzy logic variables.

Such circuits may find applications in realizing certain logic which are fuzzy in nature, such as, in making a "machine doctor".

REFERENCES

1. Zadeh, L.A., "Fuzzy Sets", Information and Control 8, 338-353 (1965).

2. Zadeh, L.A., "Fuzzy Sets and Systems", Proc. Symp. System Theory, Polytechnic Institute of Brooklyn, 29-37.

3. Zadeh, L.A., "Fuzzy Algorithms", Information and Control 12, 94-102 (1968).

4. Marinos, P.N., "Fuzzy Logic and Its Application to Switching Systems", IEEE Transactions on Computers, Vol. C-18, No. 4, April 1969.

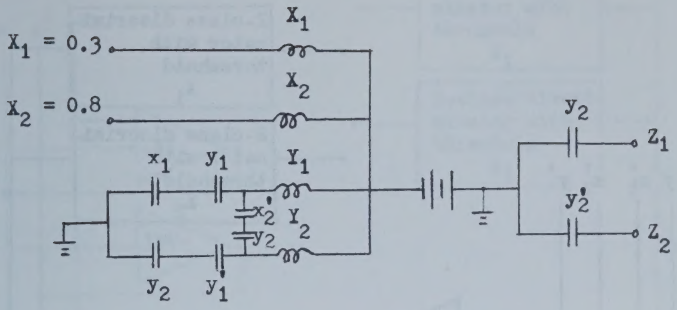


Fig. 1

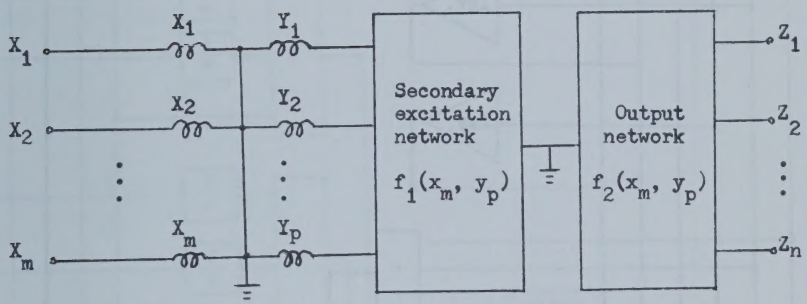


Fig. 2

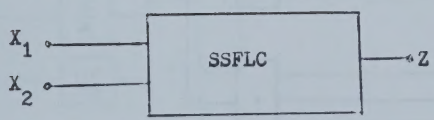


Fig. 3

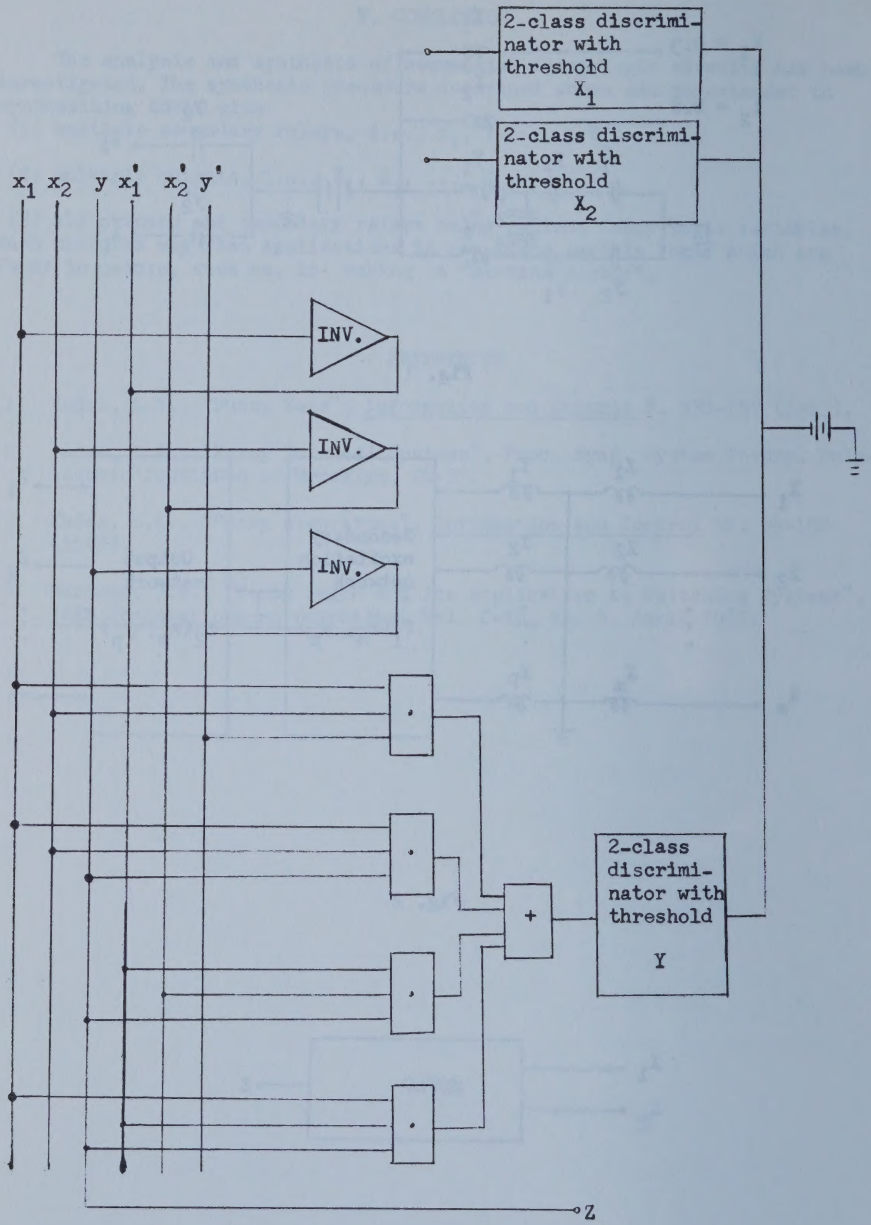

2-class discrimi-
nator with
threshold
X1
2-class discrimi-
nator with
threshold
X2
x1 x2 y x1' x2' y'
INV.
INV.
INV.
2-class
discrimi-
nator with
threshold
Y
+
Z

Fig. 4

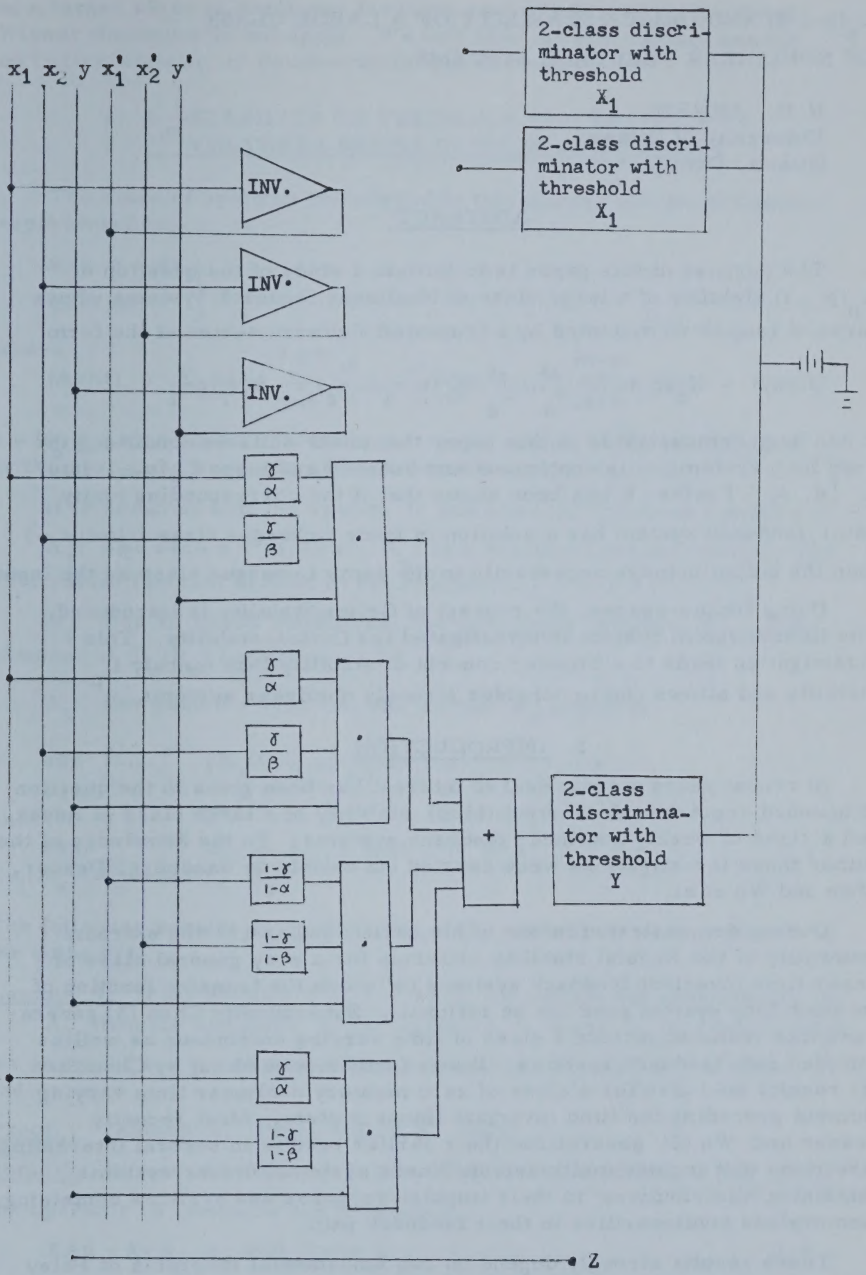

x1 x2 y x'1 x'2 y'
2-class discri-minator with threshold X1
2-class discri-minator with threshold X1
INV.
INV.
INV.
γ/α
γ/β
γ/α
γ/β
1-γ/1-α
1-γ/1-β
γ/α
1-γ/1-β
+
2-class discrimina-tor with threshold Y
Z

Fig. 5

$L_p(p \geq 1)$ AND ORLICZ STABILITY OF A LARGE CLASS OF NONLINEAR FEEDBACK SYSTEMS

N.U. AHMED
University of Ottawa
Ottawa, Canada.

ABSTRACT

The purpose of this paper is to initiate a study of the question of $L_p(p \geq 1)$ stability of a large class of Nonlinear feedback systems whose forward loop is represented by a truncated Volterra series of the form

$$(Ax)(t) = K_o(t) + \sum_{s=1}^{n<\infty} \int_o^t \cdots \int_o^t K_n(t; \tau_1 \cdots \tau_s) \prod_{i=1}^{s} x(\tau_i)\, d\tau_i .$$

It has been demonstrated in this paper that under suitable conditions the open loop system A is continuous and bounded and maps $L_p[o,\infty)$ into $L_p[o, \infty)$. Further it has been shown that if the corresponding (unity gain) feedback system has a solution in some Lebesgue class $L_p(o<p\leq \infty)$ then the output belongs necessarily to the same Lebesgue class as the input.

Using Orlicz-spaces, the concept of Orlicz Stability is introduced. The Hammerstein system is investigated for Orlicz-stability. This investigation leads to a broader concept of stability than merely L_p stability and allows one to consider strongly nonlinear systems.

1. INTRODUCTION

In recent years a great deal of interest has been given in the question of bounded-input-bounded-output (bibo) stability of a large class of linear and a class of weakly nonlinear feedback systems. To the knowledge of the author these investigations were carried out mainly by Sandberg, Desoer, Chen and Wu et al.

Desoer demonstrated in one of his earlier papers(1) the extreme generality of the Nyquist stability criterion for a very general class of linear time invariant feedback systems (of which the transfer function of the open loop system need not be rational). Subsequently Chen (3) generalized this result to include a class of time varying continuous as well as sampled data feedback systems. It was further pointed out by Chen that his results hold also for a class of zero memory nonlinear time varying element preceding the time invariant linear system. Most recently Desoer and Wu (2) generalized their earlier results in several interesting directions that include multivariable linear systems, linear systems containing 'distributions' in their impulse response and systems containing memoryless nonlinearities in their feedback path.

These results strongly depend on two fundamental theorems of Paley and Wiener as noted by the author (4).

In this paper we are interested in the question of $L_p(p \geq 1)$ stability

of a larger class of Nonlinear feedback systems for which the Paley-Wiener theorems do not apply. We will also present certain results on Orlicz stability of Hammerstein type systems.

2. L_p-STABILITY OF FEEDBACK SYSTEMS WITH VOLTERRA SERIES IN ITS FORWARD PATH

The class of systems considered in this section has the following representation:

$$\left.\begin{aligned} x &= u - Ax \\ y &= Ax \end{aligned}\right\} S$$

where

$$(Ax)(t) = K_o(t) + \sum_{s=1}^{n<\infty} \int_o^t \cdots \int_o^t K_s(t;\tau_1 .. \tau_s) \prod_{i=1}^{s} x(\tau_i) d\tau_i, \qquad 2.1$$

$t \in R_o \triangleq [o, \infty)$; and u is the input, y the output, x the error and K_o is the zero input response of the forward loop (operator A).

It is assumed that the system S satisfies the following conditions:

A_1: For each $s \in \{1, 2, \cdots n\}$, K_s is a bounded Lebesgue measurable real valued function defined in the region $o \le \tau_i < \infty$ $i = 1, 2, .. s$.

A_2: For all initial conditions, the zero input response K_o is an element of $L_p(R_o)$ for some $p \in [1, \infty]$.

A_3: For each $s = 1, 2, .. n$, the kernel K_s satisfies

$$\sup_{t \ge o} \int_o^t \cdots \int_o^t |K_s(t; \tau_1 \cdots \tau_s)| \, d\tau_1 .. d\tau_s \le c_s < \infty \qquad 2.2$$

and

$$\sup_{(\tau_1 .. \tau_s) \in R_o^s} \int_{\max(\tau_1 .. \tau_s)}^{\infty} |K_s(t; \tau_1 .. \tau_s)| \, dt \le d_s < \infty \qquad 2.3$$

The following results on the stability of the system S were reported by the author (4).

Lemma 1. If the kernels $\{K_s\}$ $s = 1, 2, .. n$ satisfy the assumptions A_1 - A_3 then for any finite n the operator A maps $L_p(R_o)$ into $L_p(R_o)$ for any $p \in [1, \infty]$ and is continuous in any ball $S_r \subset L_p(R_o)$ of finite radius r.

Proof: (Author 4, Proc. I.E.E.E. 1969)

Remark By use of lemma 1 it is easily shown that for $u, v \in S_r \subset L_p$ the operator A satisfies the inequality

$$\| Au - Av \|_p \le a(r) \| u - v \|_p \qquad 2.4$$

where $a(r) \ge o$ for $r \ge o$ and $\lim_{r \to \infty} a(r) \to \infty$.

This shows that unlike bounded linear operators, the continuity of the operator A does not imply uniform continuity throughout the space $L_p(R_o)$.

The question of L_p stability of the feedback system S is equivalent to the question of existence of a solution of the integral equation $x = u - Ax$ in L_p space. It appears to the author that additional conditions other than merely A_1 and A_3 have to be imposed on the magnitudes of the numbers $\{c_s, d_s\}$ in order to prove an existence theorem. For example, it is quite simple to show that if corresponding to the sequence of numbers $\{c_s, d_s\}$ there exists a number $r_o > o$ such that $a(r_o) < 1$ then for every $\tilde{u} \triangleq u - K_o$ in the ball $S_{r_o(1-a(r_o)} \subset L_p(R_o)$ there exists (by Banach fixed point theorem) a unique solution $x \in S_{r_o} \subset L_p(R_o)$. This result appears to be very restrictive.

A question that can be settled easily is: if there is a solution to the system S in some Lebesgue class then does it belong to the same class as u and K_o ? The answer is in the affirmative and is proved in the following proposition.

Proposition 1, If the kernels $\{K_s\}$, $s = 1, 2, .. n$ of the operator A satisfy the assumptions A_1 and A_3 and if u and K_o are elements of $L_p(R_o)$ and if the integral equation $x = u - Ax$ has a solution in some Lebesgue class $L_s(R_o)$ $(0 < s \leq \infty)$ then necessarily $y \in L_p(R_o)$ and further it is unique.

Proof (Author (4) Proc. I.E.E.E,).

It may be mentioned that the operator A as defined in the introduction is only a special case of the operators introduced by the author as Hammerstein and Uryson series in his study of the algebra of nonlinear operators on L_2 spaces (5).

In the following section we will limit ourselves to the first term of the Hammerstein series.

It will be observed that by use of Orlicz-spaces it is possible to study within this class strongly nonlinear systems and derive very weak sufficient conditions for stability.

3. ORLICZ-STABILITY OF FEEDBACK SYSTEMS WITH THE HAMMERSTEIN OPERATOR IN ITS FORWARD PATH.

In this section sufficient conditions for L_p stability as well as Orlicz-stability, of the class of systems represented by

$$\left.\begin{aligned} x &= u - \lambda Ax \\ y &= Ax \end{aligned}\right\} T$$

where for almost all $t \in R_o$

$$(Ax)(t) = (Kfx)(t) = \int_o^t K(t,\tau) f(\tau, x(\tau)) d\tau , \qquad 3.1$$

are presented under sufficiently weaker assumptions on K and f.

There are practical situations where it is not very restrictive to consider $f(t,u)$ a measurable function of t on R_o for each $u \in R$ and continuous in u on R for almost all $t \in R_o$ (Carathédory condition). But it is very restrictive to assume that f satisfies a Lipschitz condition (global) in the second variable. In fact for an appropriate class of linear operators in the forward loop it is possible to allow f to have nonlinearities with exponents $\alpha \geq 1$. This is proved in the following lemma 2.

Lemma 2. Suppose there exist real numbers $\alpha \geq 1, p_2 \geq 1$ and $\beta \geq o$ so that for almost all $t \in R_o$ $|f(t,u)| \leq g(t) + \beta |u|^{\alpha}$ with $g(t) \in L_{p_2}(R_o)$. Then the operator f maps $L_{p_1}(R_o)$ into $L_{p_2}(R_o)$ and is continuous and bounded where $p_1 = \alpha p_2$.

Proof: The first assertion of the lemma follows from the inequality $\| f(\cdot, x(\cdot)) \|_{p_2} \leq \| g \|_{p_2} + \beta (\| x \|_{p_1})^{\alpha}$ which implies that f : $L_{p_1} \rightarrow L_{p_2}$. This in turn implies continuity (6, th 2.1, pp 22) and boundedness (6, th 2.2, pp 26). This completes the proof of the lemma.

With the help of this lemma we can prove the following result. We need the following

Definition 1. An operator A mapping a Banach space B_1 into a Banach space B_2 is said to be completely continuous on $S \subset B_1$ if it is continuous on S and AS is a compact subset of B_2 (whenever S is bounded).

Proposition 2. If the kernel $K(t,\tau)$ is measurable on the triangle $o \leq \tau \leq t < \infty$ and satisfies the property that $\hat{K}(t) \triangleq (\int_o^t |K(t,\tau)|^{q_2} d\tau)^{1/q_2} \in L_{p_1}(R_o)$ (where $p_2^{-1} + q_2^{-1} = 1$) and f satisfies the hypothesis of lemma 2 then the operator $A \triangleq Kf$ maps $L_{p_1}(R_o)$ into itself and that it is completely continuous on $L_{p_1}(R_o)$.

Proof: The first part of the proposition follows from the fact that $f: L_{p_1} \rightarrow L_{p_2}$ (lemma 2) and the inequality

$$\| Ax \|_{p_1} \leq \| \hat{K} \|_{p_1} \| fx \|_{p_2} \qquad 3.2$$

for all $x \in L_{p_1}$. Thus the operator A maps $L_{p_1}(R_o)$ into itself.

For the proof of complete continuity we must prove that A is continuous (on $L_{p_1}(R_o)$) and compact (ie. maps every bounded set in L_{p_1} into a compact set in L_{p_2}).

The continuity of the operator A follows from that of the operator f and the obvious inequality $\|Ax - Ay\|_{p_1} \leq \|\hat{K}\|_{p_1} \|fx - fy\|_{p_2}$. It remains to prove the compactness. Let $D_1 \subset L_{p_1}(R_o)$ be a bounded subset. Since by lemma 2 $f: L_{p_1} \longrightarrow L_{p_2}$ is bounded $\sup_{x \in D_1} \|fx\|_{p_2} \leq d_1$ for some $0 \leq d_1 < \infty$. Define $D_2 = \{z = fx : x \in D_1\}$.

Clearly $D_2 \subset L_{p_2}$ is bounded and since for $p_2 > 1$ L_{p_2} is a reflexive Banach space D_2 is weakly compact. If $p_2 = 1$ then we, may assume that the operator f satisfies the additional property that for every $\epsilon > o$ there exists a $\delta > o$ such that $\int_E |f(t,x(t))|\, dt < \epsilon$ for all $x \in D_1$ whenever the Lebesgue measure of E ie. $\mu(E) < \delta$. Under this situation D_2 is a weakly sequentially compact subset of $L_{p_2}(R_o) = L_1(R_o)$ (9, pp 292 th 9).

Thus we may assume that D_2 is weakly (sequentially) compact. Now let $\{z_n = fx_n, x_n \in D_1\} \subset D_2$ be any sequence. It is clear that for almost every $t \in R_o$ the number

$$y_n(t) = \int_o^t K(t,\tau)\, z_n(\tau)\, d\tau$$

$$\triangleq \int_o^t K_t(\tau)\, z_n(\tau)\, d\tau \qquad 3.3$$

is defined. Further it follows from the assumptions on K that for almost every $t \in R_o$ $K_t(\tau) \in L_{q_2}[o,t)$. Therefore since D_2 is weakly compact there is a subsequence $\{z_m\}(m = n_1, n_2, \ldots)$ of $\{z_n\}$ and $z_o \in L_{p_2}(R_o)$ so that for almost all $t \in R_o$

$$y_m(t) = \int_o^t K_t(\tau)\, z_m(\tau)\, d\tau \longrightarrow \int_o^t K_t(\tau)\, z_o(\tau) d\tau \triangleq y_o(t) \qquad 3.4$$

Therefore $\lim_{m\to\infty} (y_m(t) - y_o(t)) = 0$ a.e on R_o and consequently

$$\lim_{m\to\infty} |y_m(t) - y_o(t)|^{p_1} = o \quad \text{a.e on } R_o. \qquad 3.5$$

Further $$|y_m(t) - y_o(t)| \leq \hat{K}(t)\,(d_1 + \|z_o\|_{p_2}) \qquad 3.6$$

uniformly with respect to m and for almost all $t \in R_o$. By hypothesis $\hat{K} \in L_{p_1}(R_o)$ and therefore $\{y_m - y_o\} \in L_{p_1}(R_o)$.

The function on the right hand side of the inequality 3.6 provides the dominating function required for the well known Lebesgue dominated convergence (9) theorem to hold. Therefore by application of the dominated convergence theorem we have

$$\lim_{m\to\infty} \int |y_m(t) - y_o(t)|^{p_1} dt = \int_{R_o} \lim_{m\to\infty} |y_m(t) - y_o(t)|^{p_1} dt = 0 \qquad 3.7$$

Thus $y_m \to y_o$ in the strong topology of $L_{p_1}(R_o)$ and it follows from the equality $y_o = y_o - y_m + y_m$ and the uniform boundedness of $\{z_m\} \in L_{p_2}$ and consequently that of $\{y_m\} \in L_{p_1}$ that $y_o \in L_{p_1}(R_o)$. Thus AD_1 is compact. This completes the proof.

Using Lemma 2 and proposition 2 we prove the following

Proposition 3:

Let the input $u \in L_{p_1}(R_o)$ be arbitrary, $r > o$ be a number such that $\|u\|_{p_1} < r$. Then if the feedback gain factor λ is such that $|\lambda| < \frac{r - \|u\|}{a(r)}$ where $a(r) \triangleq \sup_{x \in S_r \subset L_{p_1}} \|Ax\|_{p_1}$ then the system T has a solution $x^* \in L_{p_1}(R_o)$ and the output $y^* = Ax^* \in L_{p_1}(R_o)$.

Proof: If $|\lambda| < \frac{r - \|u\|_{p_1}}{a(r)}$ then $Bx \triangleq u - \lambda Ax$ satisfies the inequality $\|Bx\|_{p_1} \leq \|u\| + |\lambda| a(r) < r$. Therefore the operator B maps S_r into a set $T_r \subset S_r$. By proposition 2 B is a completely continuous operator acting within $L_{p_1}(R_o)$ since A is also completely continuous and $u \in L_{p_1}$. Therefore by Schauder fixed point principle (7, pp 209) there is at least one solution $x^* \in S_r$ of the system T and $y^* \in L_{p_1}(R_o)$.

Remark: It appears from the above results that nonlinearities stronger than power-nonlinearities can not be handled by use of L_p spaces. This limitation, however, can be overcome to a large extent by use of Orlicz spaces. For detailed properties of Orlicz spaces the reader is referred to Zaanen (8, pp 78), and Krasnosel'skii and Rutickii (7).

For simplicity we shall consider the Orlicz spaces L_{M_1} and L_{M_2} so that M_1, M_2 and the function N_2 complimentary to the function M_2 in the sense of Young (8, pp 76) satisfy the so called Δ_2 condition (7, pp 24) ie there exists a finite real number $\alpha' > o$ so that $M_i(2u) \leq \alpha' M_i(u)$ $i = 1, 2$ and $N_2(2u) \leq \alpha' N_2(u)$ for all $u \geq o$. We have the following result analogous to the lemma 2:

Lemma 3: Suppose there exists a real number $\beta \geq o$ and real valued measurable function $g \in L_{M_2}$ so that

$$|f(t,u)| \leq g(t) + \beta M_2^{-1}[M_1(u)]$$

and suppose the function M_2 satisfy the Δ_2 condition as defined above.

Then the operator f maps L_{M_1} into L_{M_2} and is continuous and bounded

Proof: The proof follows from Th 17.5, and Th 17.6, (Krasnoselskii and Rutickii, 7 pp 174).

Using the above lemma we obtain the following proposition analogous to proposition 2:

Proposition 4. If the Kernel $K(t,\tau)$ is measurable on the triangle $0 \le \tau \le t < \infty$ and satisfies the property that $\hat{K}(t) = \sup_{\rho_t(u, M_2) \le 1} \left| \int_o^t K(t,\tau) u(\tau) d\tau \right| \in L_{M_1}(R_o)$ (where $\rho_t(u, M_2) \triangleq \int_o^t M_2 \left| u(\tau) \right| d\tau$ and f satisfies the hypothesis of Lemma 3 then the operator $A = Kf$ maps $L_{M_1}(R_o)$ into itself and that it is completely continuous on $L_{M_1}(R_o)$.

Proof: The proof follows similar arguments as in proposition 2 and uses the fact that L_{M_2} is a reflexive Banach space (8, pp 158, th 6) since by our basic assumption both, M_2 and N_2 satisfy Δ_2 condition. Therefore a bounded subset of L_{M_2} is weakly compact. The notable difference in the proof is that the expressions 3.6 and 3.7 are replaced by

$$\text{(i)} \qquad \left| y_m(t) - y_o(t) \right| \le \hat{K}(t) \left(d_1 + \| z_o \|_{M_2} \right) \qquad 3.6'$$

where $d_1 = \sup_{x \in D_1} \| fx \|_{M_2}$ and

$$\hat{K}(t) = \sup_{\rho_t(u, M_2)} \left| \int_o^t K(t,\tau)\, u(\tau) dt \right| \in L_{M_1}$$

and (ii) $$\lim \| y_m - y_o \|_{M_1} \le \sup_{\rho(v, N_1) \le 1} \int_{R_o} \lim_m \left| y_m(t) - y_o(t) \right| \left| v(t) \right| dt = 0 \qquad 3.7'$$

The last inequality follows from 3.6' and the Lebesgue dominated convergence theorem. This completes the sketch of the proof.

Combining lemma 3, proposition 4 and Schauder fixed point theorem we obtain the following result.

Proposition 5. Let the input $u \in L_{M_1}(R_o)$ be arbitrary and let $r > o$ be a finite real number such that $\| u \|_{M_1} < r$. Then if the feedback gain factor λ is such that $|\lambda| < \dfrac{r - \|u\|_{M_1}}{a(r)}$ where $a(r) = \sup_{x \in S_r \subset L_{M_1}} \| Ax \|$ then the system T has a solution $x^* \in L_{M_1}(R_o)$ and the output $y^* = Ax^* \in L_{M_1}(R_o)$.

CONCLUSION

In this paper we presented some results on the L_p $(p \ge 1)$ stability of a large class of Nonlinear feedback systems with Volterra series in the forward loop.

Sufficient conditions for L_p $(p \ge 1)$ stability of nonlinear feedback

systems with Hammerstein operator in the forward loop were presented. Lipschitz conditions were not used and nonlinearities with exponents $a > 1$ were allowed. Orlicz-stability was introduced and the Hammerstein system with nonlinearities much stronger than power-nonlinearities was investigated for Orlicz stability. Exponential nonlinearities arising in Stiction phenomenon may be handled by use of Orlicz spaces provided Δ_3 and Δ^2 conditions (7, pp 35, 46) are allowed for the functions M_1 and M_2 that appear in the Orlicz spaces L_{M_1} and L_{M_2}.

REFERENCES

(1) C.A. Desoer "A general formulation of the Nyquist criterion" I.E.E.E. Trans. circuit theory vol CT-12-pp 230-234 June 1965.

(2) C.A. Desoer and M.Y. Wu "Stability of mulitple loop feedback linear time invariant systems", J. Math. Anal. and Appl. vol 23, July 1968.

(3) C.T. Chen "L_p stability of linear time varying feedback systems", SIAM J. Control, vol 6 pp 186-193 May 1968.

(4) N.U. Ahmed, " On $L_p(b \geq 1)$ stability of a class of Nonlinear systems" (To appear in Proc. I.E.E.E., 1969).

(5) N.U. Ahmed, "An Algebra of Nonlinear Operators on L_2 Spaces", Pakistan Jr of Sc. and Ind. Research vol 11, No 4, Oct. 1968.

(6) M.A. Krasnoselskii, "Topological methods in the theory of Nonlinear Integral equations" Macmillan Co. N.Y., 1964 pp 20-29.

(7) M.A. Krasnoselskii and Y . B. Rutickii, " Convex functions and Orlicz spaces" Noordhoff Ltd., Groningen, the Netherlands, 1961, pp. 174-75.

(8) A.C. Zaanen, "Linear analysis" P. Noordhoff - Groningen, 1964, pp 78, pp 154.

(9) N. Dunford and J.T. Schwartz, "Linear operators, part 1 ", Interscience, 1964, pp 292.

---

* This work was supported in part by the National Research Council of Canada under Grant No. A-7109.

STABILITY OF NONLINEAR RANDOMLY SAMPLED SYSTEMS†

R. G. AGNIEL and E. I. JURY
Department of Electrical Engineering and Computer Sciences,
Electronics Research Laboratory, University of California,
Berkeley, California 94720

INTRODUCTION

Recently, more and more attention has been devoted to sampled-data systems with a random sampling period. The sampling period may be random for physical reasons, when it is contaminated by noise because of cheap equipment ("jitter") or when the sampling is governed by a stochastic process like a radar or sonar echo ("skip" process). The sampling period is sometimes let to be random intentionally, for instance when a time-shared computer governs several plants or to avoid jamming in communication lines. Moreover, keeping the sampling period exactly constant is only of theoretical interest and requires expensive equipment.

Lately, randomly sampled systems appeared to be of great interest in biocontrol systems. The latest models for eye tracking movements [1] as well as for human models [2] appear to be randomly sampled multinonlinear systems. Since most sampled-data control systems are also inherently nonlinear (quantization), this work which deals with nonlinear randomly sampled systems seems to be timely and warranted.

Linear randomly sampled systems have already been extensively studied in the literature by Kalman [3,4], Bergen [5,6] and Bharucha [7]. Moreover, discrete systems with piecewise constant parameters on random intervals of time have also been investigated by Bellman [8], Bertram and Sarachik [9], Bucy [10], Kushner [11] and many others.

Using the latest results by Bucy [10] on stochastic Lyapunov functions and supermartingales, this work provides a method of study of saturating type randomly sampled systems.

I. STATEMENT OF THE PROBLEM

a. Description of the System

Consider the system depicted in figure 1 satisfying the following assumptions:

1. The sampling process is constituted of independent identically distributed successive sampling intervals. The successive sampling intervals will be denoted $T_0, T_1, T_2, \ldots$ and the successive sampling instants $t_0, t_1, t_2, \ldots$ are indicated in figure 2. Obviously, the sequence $\{T_i\}$ is stationary and the expectation of any function of T_i is independent of i.

2. The system considered has a null input, i.e.,

$$r(t) = 0 \quad \text{for all } t \geq 0. \tag{1}$$

3. The sampling and hold element is such that, if y(t) is its input and $\sigma(t)$ its output, we have

$$\sigma(t) = \sigma(t_k) \triangleq \sigma_k = y(t_k) \quad \text{for } t_k \leq t < t_{k+1}. \tag{2}$$

4. The nonlinearity is continuous, of sector type and with bounded slope, i.e., if σ is the input of the nonlinearity and $m(\sigma)$ its output, we have

$$0 \leq \sigma m(\sigma) \leq k_0 \sigma^2, \quad \text{for all } \sigma \tag{3}$$

$$\frac{dm}{d\sigma} \leq K, \quad \text{for all } \sigma \tag{4}$$

The constants k_0 and K can be found by

$$k_0 = \sup_{\sigma} \frac{m(\sigma)}{\sigma} \tag{5}$$

$$K = \sup_{\sigma} \frac{dm}{d\sigma} \tag{6}$$

Obviously, $k_0 \leq K$ if $m(0) = 0$. For a saturating nonlinearity, the maximum slope is for $\sigma = 0$ and $k_0 = K$.

5. The plant is linear and time invariant; it is described by the equation

$$\dot{\underset{\sim}{X}} = A\underset{\sim}{X} + \underset{\sim}{b}m, \tag{7}$$

where $\underset{\sim}{X}$ is the n-dimensional state vector,

A is an $n \times n$ constant matrix,

$\underset{\sim}{b}$ is an n-dimensional constant vector,

m is the output of the nonlinearity.

Since $m(t) = m(t_k) \triangleq m(k)$ for $t_k \leq t < t_{k+1}$, we have (see, for instance, Kalman and Bertram [12])

$$\underset{\sim}{X}(k+1) = G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m(k), \tag{8}$$

where $T_k = t_{k+1} - t_k$ is the k^{th} sampling interval, $\underset{\sim}{X}(k)$ is the state vector at time t_k,

$$G(k) \triangleq G(T_k) = e^{AT_k}, \tag{9}$$

$$\underset{\sim}{h}(k) \triangleq \underset{\sim}{h}(T_k) = \int_0^{T_k} e^{Az}\underset{\sim}{b}\, dz. \tag{10}$$

Moreover, the eigenvalues of the matrix $E[G(k) \otimes G(k)]$ lie inside the unit circle, where E denotes the expectation of the matrix $G(k) \otimes G(k)$,* and $G(k) \otimes G(k)$ is the first Krönecker product of $G(k)$ by itself (see Kalman [3]). In symbols, we write

$$\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| < 1. \tag{11}$$

Evidently, one can pick any representation for the linear plant. However, the canonical form seems to be the most desirable one since it will reduce the complexity of the forthcoming computations.

6. The feedback is a linear combination of the states, i.e.,

$$y(t) = \underset{\sim}{a}'X(t), \quad \underset{\sim}{a} \text{ being a constant vector.}$$

The equation of the system is

$$\underset{\sim}{X}(k+1) = G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m[\underset{\sim}{a}'\underset{\sim}{X}(k)]. \tag{12}$$

b. Stochastic Stability

Given an initial condition $\underset{\sim}{X}_0$ at time t_{n_0}, there exists as many different outputs as there are different sequences $\{T_i\}$. The convergence properties of the stochastic process generated by the output form the stochastic stability of the system. The following definitions are the stochastic counterpart of the deterministic Lyapunov stability concepts; the terminology is borrowed from Kozin [13,14,15].

Let $X(n,X_0,n_0)$ denote the state vector at time t_n when the initial state was $\underset{\sim}{X}_0$ at time t_{n_0}.

1. Almost sure stability: The equilibrium solution $\underset{\sim}{X}_e$ is said to be almost surely stable if, given $\varepsilon,\varepsilon' > 0$, there exists $\delta(\varepsilon,\varepsilon',n_0)$ such that $\|\underset{\sim}{X}_0 - \underset{\sim}{X}_e\| < \delta$ implies

$$P[\sup_{n \geq n_0} \|\underset{\sim}{X}(n,\underset{\sim}{X}_0,n_0) - \underset{\sim}{X}_e\| > \varepsilon'] < \varepsilon$$

*The expectation of a matrix is the matrix formed by the expectation of all its elements.

where P denotes the probability and $\|\cdot\|$ any vector norm.

2. Almost sure asymptotic stability. The equilibrium solution $\underset{\sim}{X}_e$ is said to be almost surely asymptotically stable if:

--it is almost surely stable,

--for any $\varepsilon > 0$, there exists δ' such that $\|\underset{\sim}{X}_0 - \underset{\sim}{X}_e\| < \delta'$ implies

$$\lim_{\delta \to \infty} P\left[\sup_{n \geq \delta} \|\underset{\sim}{X}(n, \underset{\sim}{X}_0, n_0) - \underset{\sim}{X}_e\| > \varepsilon\right] = 0$$

i.e., $\|\underset{\sim}{X}(n, \underset{\sim}{X}_0, n_0) - \underset{\sim}{X}_e\| \xrightarrow{a.s.} 0$

3. Stability of the m^{th} moment: The equilibrium solution $\underset{\sim}{X}_e$ is said to possess stability of the m^{th} moment if, given $\varepsilon > 0$ there exists $\delta(\varepsilon, n_0)$ such that $\|\underset{\sim}{X}_0 - \underset{\sim}{X}_e\| < \delta$ implies

$$\sup_{n \geq n_0} E[\|\underset{\sim}{X}(n, \underset{\sim}{X}_0, n_0)\|^m] < \varepsilon$$

4. Asymptotic stability of the m^{th} moment: The equilibrium solution $\underset{\sim}{X}_e$ shows asymptotic stability of the m^{th} moment if

--it possesses stability of the m^{th} moment if

--there exists $\delta' > 0$ such that $\|\underset{\sim}{X}_0 - \underset{\sim}{X}_e\| < \delta'$ implies

$$\lim_{n \to \infty} E[\|\underset{\sim}{X}(n, \underset{\sim}{X}_0, n_0) - \underset{\sim}{X}_e\|^m] = 0$$

Evidently, these stability concepts are the stochastic extension of the deterministic Lyapunov stability concepts; but there are as many extensions as there are modes of convergence. Here, the extension has been done for convergence of the m^{th} moment and almost sure convergence.

Almost sure asymptotic stability seems to be the most desirable mode of stability since for every experiment the output converges to its equilibrium.

c. Objective of this Work

Given the system described in section I.a, this work gives a computational procedure to derive the maximum k_0 for almost sure asymptotic stability, i.e., to find the largest possible sector for the nonlinearity.

We will see later that this method provides the largest sector for almost sure asymptotic stability and asymptotic stability of the second moment.

The same system with periodic sampling was considered by Pearson and Gibson [16], using the Lurê form and Jury and Lee [17] using the frequency method.

II. PRELIMINARIES

a. Stochastic Lyapunov functions

Consider the random system described by

$$\begin{cases} \underset{\sim}{X}(k+1) = f[\underset{\sim}{X}(k), T_k] \\ \underset{\sim}{X}(0) = \underset{\sim}{X}_0 \end{cases} \tag{13}$$

where $\underset{\sim}{X}(k)$ is the n-dimensional random state vector at time t_k,

$\{T_k\}$ is a random sequence,

$f(\cdot)$ is a continuous vector function,

$\underset{\sim}{X}_0$ is the initial state at time t_0.

Define an equilibrium solution $\underset{\sim}{X}_e$ by

$$\underset{\sim}{X}_e = f[\underset{\sim}{X}_e, T_n], \quad \text{for all } n \tag{14}$$

It will not reduce the generality of the treatment to take $\underset{\sim}{X}_e = 0$, i.e., to change $\underset{\sim}{X}$ into $\underset{\sim}{X} - \underset{\sim}{X}_e$.

The two following lemmas, due to Bucy [10], express the intuitive idea

that, if $V_n \triangleq V(\underset{\sim}{X}(n))$ is a positive function representing the energy of the system in some way, then

--if EV_n is nonincreasing with n, we have stability

--if EV_n is decreasing with n, we have asymptotic stability.

Lemma 1 (Bucy [10]): Almost sure stability

Consider the system described by (13) with $\|\underset{\sim}{X}_0\| < \infty$ and 0 being an equilibrium solution. $\underset{\sim}{X}(n)$ will denote the solution of (13) with $\underset{\sim}{X}(0) = \underset{\sim}{X}_0$. Suppose there exists a positive definite continuous function $V(\underset{\sim}{X})$ called stochastic Lyapunov function such that

i. $V(\underset{\sim}{0}) = 0$,

ii. $V(\underset{\sim}{X})$ is finite if $\|\underset{\sim}{X}\|$ is finite,

iii. $V(\underset{\sim}{X}) \to \infty$ as $\|\underset{\sim}{X}\| \to \infty$.

Let $\mathcal{B}_n$ be the minimal σ-field over $\{\underset{\sim}{X}_0, \underset{\sim}{X}(1), \ldots, \underset{\sim}{X}(n)\}$; then if $V_n \triangleq V(\underset{\sim}{X}(n))$ is a supermartingale sequence, the origin is almost surely stable. With the definitions and notations of Loève [19] (Chapter VII, § 24 and 25, Chapter VIII, § 23), this condition is written

$$E^{\mathcal{B}_n} V_{n+1} - V_n \leq 0 \quad \text{a.s.} \tag{15}$$

Lemma 2 (Bucy [10]): Almost sure asymptotic stability

Suppose that the assumptions of Lemma 1 hold, i.e., we have almost sure stability of the origin and, moreover, there exists a continuous function $\gamma(\cdot)$ such that

i. $\gamma(0) = 0$,

ii. $\gamma(z) = 0$ implies $z = 0$,

iii. $\gamma(z) > 0$ whenever $z \neq 0$,

iv. $E^{\mathcal{B}_n} V_{n+1} - V_n \leq -\gamma(\|\underset{\sim}{X}(n)\|)$ a.s. (16)

Then, the origin is almost surely asymptotically stable and, moreover, $EV_n \to 0$.

b. Application to Linear Systems

Consider the linear plant of the system described in Section I.a. With a zero input, this open loop linear system is described by [from (8), (9) and (10)]

$$\underset{\sim}{X}(k+1) = G(k)\underset{\sim}{X}(k). \tag{17}$$

Taking the stochastic Lyapunov function of the form

$$V_n = \underset{\sim}{X}'(n)B\underset{\sim}{X}(n), \quad B \text{ positive definite matrix}, \quad \underset{\sim}{X}' = \text{Transpose of } \underset{\sim}{X} \tag{18}$$

it is straightforward to find that the system is almost surely asymptotically stable and also shows asymptotic stability of the second moment if there exists constant positive definite matrices B and Q such that

$$-Q = E[G'(k)BG(k)] - B \tag{19}$$

On the other hand, it is well known (see Kalman [3,4], Bharucha [7]) that the system of the form (17) shows asymptotic stability of the second moment if and only if

$$\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| < 1 \tag{20}$$

We will see in the next section that (20) is a necessary and sufficient condition for (19) to have a unique solution.

c. Solution of $-Q = E[G'(k)BG(k)] - B$

Problem: Given a positive definite matrix Q, under what conditions can we find a positive definite matrix B satisfying equation (19)?

--Existence and uniqueness: Although G(k) is not necessarily symmetric, G'(k)BG(k) is symmetric and it does not affect the generality of the proof to

assume B and Q symmetric (see Gantmacher [18], p. 294). If b_{ij} and q_{ij} are the general elements of B and Q respectively, equation (19) can be rewritten as

$$(E[G(k) \otimes G(k)]-I) \begin{bmatrix} b_{11} \\ b_{12} \\ \vdots \\ b_{1n} \\ b_{22} \\ b_{23} \\ \vdots \\ b_{2n} \\ b_{33} \\ \vdots \\ b_{nn} \end{bmatrix} = - \begin{bmatrix} q_{11} \\ q_{12} \\ \vdots \\ q_{1n} \\ q_{22} \\ q_{23} \\ \vdots \\ q_{2n} \\ q_{33} \\ \vdots \\ q_{nn} \end{bmatrix} \tag{21}$$

where $G(k) \otimes G(k)$ is the first Krönecker product of G(k) by itself. From (21) we see that given the right-hand side vector formed with the elements of Q, we can determine the vector formed with the elements of B, provided

$$\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| \neq 1.$$

Since B and Q are taken symmetric, we can determine B completely from (21).

--Condition for a positive definite B

• Since condition (19) implies that the system described by (17) shows stability of the second moment, it is necessary that (20) be verified (see Kalman [3], Bharucha [7]) and the, if $\underset{\sim}{X}'(k) = [x_1(k), x_2(k), \ldots, x_n(k)]$,

$$E[x_i(k)x_j(k)] \to 0 \text{ as } k \to \infty \quad \text{for all } i,j = 1,2,\ldots,n \tag{22}$$

• Conversely, suppose that B is not positive definite; then there exists a state $\underset{\sim}{X}_0$ such that $\underset{\sim}{X}_0'B\underset{\sim}{X}_0 < 0$. This implies that

$$E[\underset{\sim}{X}'(1)B\underset{\sim}{X}(1)] = \underset{\sim}{X}_0'B\underset{\sim}{X}_0 - \underset{\sim}{X}_0'Q\underset{\sim}{X}_0 < 0 \tag{23}$$

since Q is positive definite.

By induction, $E[\underset{\sim}{X}'(k)B\underset{\sim}{X}(k)] < 0$ for all $k \geq 0$, which contradicts (22). Therefore B is positive definite.

Conclusion: Given a positive definite matrix Q, there exists a unique positive definite matrix B satisfying (19), provided

$$\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| < 1,$$

i.e., if the open loop linear system described by (17) shows asymptotic stability of the second moment.

III. ANALYSIS OF THE NONLINEAR SYSTEM

a. Choice of a Lyapunov Function

Take the Lyapunov function in the Luré form

$$V_k \triangleq V(\underset{\sim}{X}(k)) = \underset{\sim}{X}'(k)B\underset{\sim}{X}(k) + \alpha\int_0^{\sigma(k)} m(\phi)d\phi, \quad \alpha \geq 0 \tag{24}$$

where α is a positive scalar and B an $n \times n$ positive definite symmetric matrix. Clearly, V_k is positive definite. Moreover, it has been verified extensively in the literature that V_k satisfies the conditions i., ii., and iii. of Lemma 1.

b. Derivation of the Stability Condition

Following the same argument as that in Pearson and Gibson [16], and since the nonlinearity is continuous, the application of MacLaurin's formula yields

$$m(\sigma_k+h) = m(k) + h\,\frac{dm}{d\sigma}(\theta), \quad \sigma_k \leq \theta \leq \sigma_k + h \tag{25}$$

which implies

$$\int_{\sigma_k}^{\sigma_{k+1}} m(\phi)\,d\phi = m(k)\,[\sigma_{k+1}-\sigma_k] + \frac{(\sigma_{k+1}-\sigma_k)^2}{2}\,\frac{dm}{d\sigma}(\Theta) \tag{26}$$

and from (4)

$$\int_{\sigma_k}^{\sigma_{k+1}} m(\phi)\,d\phi \leq m(k)[\sigma_{k+1}-\sigma_k] + \frac{K}{2}\,(\sigma_{k+1}-\sigma_k)^2 \tag{27}$$

Consider

$$V_k = E^{\mathcal{B}_k}V_{k+1} - V_k = E^{\mathcal{B}_k}\left[\underset{\sim}{X}'(k+1)B\underset{\sim}{X}(k+1)+\alpha\int_0^{k+1} m(\phi)\,d\phi\right] - \underset{\sim}{X}'(k)B\underset{\sim}{X}(k) - \alpha\int_0^{k} m(\phi)\,d\phi. \tag{28}$$

Since V_k is $\mathcal{B}_k$-measurable, (28) can be rewritten as (see Loève [19], § 25-2-3, p. 350)

$$\Delta V_k = E^{\mathcal{B}_k}\left[\underset{\sim}{X}'(k+1)B\underset{\sim}{X}(k+1) - \underset{\sim}{X}'(k)B\underset{\sim}{X}(k) + \alpha\int_{\sigma_k}^{\sigma_{k+1}} m(\phi)\,d\phi\right] \tag{29}$$

and using the order preservation property of conditional expectations (see Loeve [19], §25-1-1, p. 347) and equation (27).

$$\Delta V_k \leq E^{\mathcal{B}_k}[\underset{\sim}{X}'(k+1)B\underset{\sim}{X}(k+1) - \underset{\sim}{X}'(k)B\underset{\sim}{X}(k) + \alpha m(k)(\sigma_{k+1}-\sigma_k) + \frac{\sigma K}{2}\,(\sigma_{k+1}-\sigma_k)^2] \tag{30}$$

Using (12), and since the T_k's (and then the G(k)'s and h(k)'s) are independent, (30) can be rewritten as (see Loève [19], 25-2-2, p.249)

$$\begin{aligned}\Delta V_k \leq E\{&[G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m(k)]'B[G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m(k)] - \underset{\sim}{X}'(k)B\underset{\sim}{X}(k)\\ &+ \alpha m(k)[\underset{\sim}{a}'(G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m(k)) - \underset{\sim}{a}'\underset{\sim}{X}(k)]\\ &+ \frac{\alpha K}{2}\,[\underset{\sim}{a}'(G(k)\underset{\sim}{X}(k) + \underset{\sim}{h}(k)m(k)-\underset{\sim}{X}(k))]^2\}\end{aligned} \tag{31}$$

the expectation being taken with respect to T_k in the usual sense.

After expanding and rearranging, and by addition of the null quantity $\underset{\sim}{a}'\underset{\sim}{X}(k)m(k) - \frac{\sigma_k}{m(k)}\,m^2(k)$, (31) can be rewritten as

$$\Delta V_k \leq -\underset{\sim}{X}'(k)C\underset{\sim}{X}(k) + 2u'\underset{\sim}{X}(k)m(k) - rm^2(k) \tag{32}$$

where C is an n × n matrix defined by

$$C = -E[G'(k)BG(k)] + B - \frac{\alpha K}{2}\,E[(G(k)-I)'\underset{\sim\sim}{a}\underset{\sim\sim}{a}'(G(k)-I)], \tag{33}$$

u' is an n-dimensional row vector defined by

$$u' = E[\underset{\sim}{h}'(k)BG(k)] + \frac{\alpha K}{2}\,E[\underset{\sim}{h}'(k)\underset{\sim\sim}{a}\underset{\sim\sim}{a}'(G(k)-I)] - \frac{\alpha}{2}\,E[a'(G(k)-I] + \frac{\underset{\sim}{a}'}{2}, \tag{34}$$

r is a scalar defined by

$$\begin{aligned}r = -E[\underset{\sim}{h}'(k)B\underset{\sim}{h}(k)] - \alpha E[\underset{\sim}{a}'h(k)]\\ - \frac{\alpha K}{2}\,E[h'(k)\underset{\sim}{a}]^2 + \frac{\sigma_k}{m(k)},\end{aligned} \tag{35}$$

and (32) can be rewritten as

$$\Delta V_k \leq -\underset{\sim}{Z}'(k)\Phi\underset{\sim}{Z}(k) \tag{36}$$

where

† m(k) is also $m(\sigma_k)$ at time t_k

$$Z(k) = \begin{vmatrix} \underset{\sim}{X}(k) \\ m(k) \end{vmatrix} \quad \text{and} \quad \Phi = \begin{vmatrix} C & -u \\ -u & r \end{vmatrix} \tag{37}$$

and V_k will satisfy all the hypotheses of lemma 2 if Φ is positive definite, i.e., (see Gantmacher [18], p. 46 and p. 306),

$$\begin{cases} C \text{ is positive definite} \\ r > \underset{\sim}{u}'C^{-1}\underset{\sim}{u} \end{cases} \tag{38}$$

and then we will have almost sure asymptotic stability and asymptotic stability of the second moment.

IV. APPLICATION OF THE RESULTS TO PRACTICAL COMPUTATIONS

a. Feasibility of the Computations

In practical situations, we are interested in finding the maximum k_0 for almost sure asymptotic stability. We have to determine a positive definite B and an $\alpha > 0$ so that k_0 given by (38) will be maximum. The general idea is to leave α as a free parameter until the final minimization.

Take any positive definite C; (33) implies that

$$-\left[C + \frac{\alpha K}{2} E\{(G(k)-I)'\underset{\sim\sim}{a}\underset{\sim\sim}{a}'(G(k)-I)\}\right] = E[G'(k)BG(k)]-B. \tag{39}$$

Obviously,

$$Q(\alpha) = C + \frac{\alpha K}{2} E\{(G(k)-I)'\underset{\sim\sim}{a}\underset{\sim\sim}{a}'(G(k)-I)\} \tag{40}$$

is positive definite; therefore we have to solve an equation of the form

$$-Q = E[G'(k)BG(k)] - B, \quad Q \text{ positive definite}$$

and we already proved in section II.c, that we can always find a positive definite B provided $\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| < 1$, which is an assumption.

Thus for any $\alpha > 0$, the computation of B is always possible.

b. Practical Computations

In practice, the computations are done as indicated on the flow chart (figure 3).

From (38) and since $k_0 \leq \frac{m(k)}{\sigma_k}$, k_0 is given by

$$k_0 = \left[\min_{\alpha \geq 0}\{E[\underset{\sim}{h}'(k)B(\alpha)\underset{\sim}{h}(k)] + \alpha E[\underset{\sim}{a}'\underset{\sim}{h}(k)] + \frac{\alpha K}{2} E[h'(k)a]^2 + u'(\alpha)C^{-1}u(\alpha)\}\right]^{-1}. \tag{41}$$

Remarks

1) The essential point in the stability of the periodically sampled case was for the linear plant to be asymptotically stable, i.e., $|\lambda_i\{G(T)\}| < 1$; here, we insure almost sure asymptotic stability of the linear plant by

$$|\lambda_i(EG(T_k) \otimes G(T_k))]\} | < 1$$

2) The optimization performed in (41) is always very simple since we have to minimize a polynomial of degree two in α with respect to α; moreover, there is no difficult matrix inversion.

3) For $\alpha = 0$, we have the usual quadratic form as a Lyapunov function. By setting $\alpha = 0$ in (41), we obtain the best result possible with an optimized quadratic form. The results given by the Lurė form are usually better than those given by a quadratic form alone (see the remark of the illustrative example).

V. ILLUSTRATIVE EXAMPLE

Consider the system depicted in figure 1 with the following transfer function:

$$G(s) = \frac{2.107s + 5.977}{(s+1)(s+2)}$$

In canonical form, the system can be written as (with the notations previously introduced)

$$\underset{\sim}{X}(k+1) = \begin{bmatrix} e^{-T_k} & 0 \\ 0 & e^{-2T_k} \end{bmatrix} \underset{\sim}{X}(k) + \begin{bmatrix} 1-e^{-T_k} \\ \frac{1-e^{-2T_k}}{2} \end{bmatrix} m(k)$$

$\sigma_k = |-3.87 \quad 1.763|\ \underset{\sim}{X}(k)$, i.e.,

$$G(k)^{\dagger} = \begin{bmatrix} e^{-T_k} & 0 \\ 0 & e^{-2T_k} \end{bmatrix},\ h(k) = \begin{bmatrix} 1-e^{-T_k} \\ \frac{1-e^{-2T_k}}{2} \end{bmatrix},\ \underset{\sim}{a} = \begin{bmatrix} -3.87 \\ 1.763 \end{bmatrix}$$

Nonlinearity: The slope is taken less than one. For a saturating nonlinearity, we will have to verify that $k_0 < K = 1$.

Sampling process: The T_k's are independent and identically Poisson distributed with parameter β, i.e., the probability of having a sampling point during the interval of time dt is β dt, β independent of t. It is well known (see Loève [19]) that this corresponds to a period T_k with density $f(\kappa) = \beta e^{-\beta\kappa}$.

Moreover, $F(s) = E(e^{-sT_k}) = \frac{\beta}{\beta+s}$. In our case, take $\beta = 1.4$; then

$$F(s) = E[e^{-sT_k}]\ \frac{1.4}{s+1.4}$$

and

$$F(1) = 0.58333 \qquad F(2) = 0.41176$$
$$F(3) = 0.31818 \qquad F(4) = 0.25929$$

Computations: We will follow the flow chart as indicated on figure 3.

a) Choice of C:

$$\text{Take} \qquad C = C^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

b) Computation of Q(α):

$$E[(G(k)-I)'\underset{\sim}{a}\underset{\sim}{a}'(G(k)-I)] = \begin{bmatrix} 14.9769[F(2)-2F(1)+1], & -6.8228[1-F(1)-F(2)+F(3)] \\ -6.8228[1-F(1)-F(2)+F(3)], & 3.1081[1-2F(2)+F(4)] \end{bmatrix}$$

$$Q(\alpha) = C + \frac{K}{2} E(G(k)-I)'\underset{\sim}{a}\underset{\sim}{a}'(G(k)-I)] = \begin{bmatrix} 1+1.83537\alpha, & -1.10218\alpha \\ -1.10218\alpha, & 1+0.67716\alpha \end{bmatrix}$$

c) Computation of B(α):

$$-Q(\alpha) = E[G'(k)BG(k)] - B$$

$$-Q(\alpha) = \begin{bmatrix} b_{11}[F(2)-1] & b_{12}[F(3)-1] \\ b_{12}[F(3)-1] & b_{22}[F(4)-1] \end{bmatrix}$$

$$B(\alpha) = \begin{bmatrix} 1.69998 + 2.12011\alpha, & -1.61652\alpha \\ -1.61652\alpha & 1.34998 + 0.91416\alpha \end{bmatrix}$$

d) Computation of k_0: We have

$$k_0 = [\min_{\alpha \geq 0} \{E(\underset{\sim}{h}'(k)B(\alpha)\underset{\sim}{h}(k)) + E(\underset{\sim}{a}'\underset{\sim}{h}(k))^2 + \frac{\alpha K}{2} E(\underset{\sim}{h}'(k)\underset{\sim}{a}) + u'C^{-1}u\}]^{-1}$$

with

†In this example $|\lambda_i\{G(T_k) \otimes G(T_k)\}| < 1$, by computing the actual values.

$$u' = E[\underset{\sim}{h}'(k)B(\alpha)G(k) + \frac{\alpha K}{2} E[\underset{\sim}{h}'(k)\underset{\sim\sim}{aa}'(G(k)-I)] - \frac{\alpha}{2} E[\underset{\sim}{a}'(G(k)-I)] + \frac{\underset{\sim}{a}'}{2}$$

$$E[\underset{\sim}{h}'(k)B(\alpha)\underset{\sim}{h}(k)] = 0.56371 + 0.34202\alpha$$

$$E[\underset{\sim}{a}'\underset{\sim}{h}(k)] = -0.36803$$

$$1/2\ E[(\underset{\sim}{h}'(k)\underset{\sim}{a}')^2] = 0.70288$$

$$E[\underset{\sim}{h}'(k)B(\alpha)G(k)] = |0.29167 + 0.312\ ,\ 0.10294 - 0.08156\alpha|$$

$$1/2\ E[\underset{\sim}{h}'(k)\underset{\sim\sim}{aa}'(G(k)-I)] = |-1.28432,\ 0.76360|$$

$$E[\underset{\sim}{a}'(G(k)-I)] = |0.80625,\ -0.51853|$$

Finally we have

$$k_0 = [\min_{\alpha \geq 0} 4.57276\alpha^2 - 4.83987\alpha + 6.4909]^{-1}.$$

The minimum is when $\alpha = 0.5292$, and then $k_0 = 0.192$ (note that $k_0 < 1$).

Remark

If we had considered an optimized quadratic form, i.e., $\alpha = 0$, we would have obtained

$$k_0 = 1/6.4909 = 0.154.$$

In this particular case, the choice of the Lure form as stochastic Lyapunov function gives a stability sector 25% larger than a quadratic form alone.

VI. CONCLUSION

This paper provides a computational procedure to determine the largest stability sector of the nonlinearity for almost sure asymptotic stability. Although this procedure is optimized, one can expect to find the largest sector by providing a better optimization.

This method can also be easily extended to multiple-input--multiple-output multi-nonlinear systems, but the computations become more complicated.

One can also derive a sufficient stability condition when the sampling process is constituted by a Markov chain with a finite number of states. One can show that a stability condition can be obtained by considering only the states contained in the ergodic set (see Agniel [20]).

Finally, extension to different types of nonlinearity is theoretically possible; however, this extension has a number of pitfalls in practice. In particular, consideration of nonlinearities in the sector $(0,\infty)$ yields poor results (or no results). For relay type nonlinearities, the system usually shows random oscillations and of course it is not almost surely asymptotically stable.

REFERENCES

[1] J.D. Forster, "A stochastic revised sampled-data model for eye tracking movement," S.M. Thesis, Report MVT-68-2, Man-vehicle Laboratory, Center for Space Research, Mass. Inst. Tech., Cambridge, Mass.

[2] G. A. Bekey, J. M. Biddle and A. J. Jacobson, "The effect of a random sampling interval on a sampled-data model of the human operator," Proc. Third Conf. on Manual Control, Report NASA-SP-144, 1967.

[3] R. E. Kalman, "Analysis and synthesis of linear dynamical systems operating on randomly sampled data," Ph.D. Dissertation, Columbia University, 1957.

[4] _____, "Control of randomly varying linear dynamical systems," Proc. Symp. Appl. Math., Vol. 43, pp. 287-298, 1961.

[5] A. R. Bergen, "Stability of systems with randomly time varying parameters," IRE Trans. Automatic Control, Vol. AC-5, pp. 265-269,

September 1969.

[6] _____, "On the statistical design of linear random sampling systems," Proc. First Congress IFAC, Butterworth Scientific Publications, London, Vol. 1, pp. 162-167, 1961.

[7] B. H. Bharucha, "On the stability of randomly varying systems," Ph.D. Dissertation, Univ. of California, Berkeley, 1961.

[8] R. Bellman, "Lomit theorems for non-commutative operations," Duke Math. J., Vol. 21, pp. 491-500, 1954.

[9] J. E. Bertram and P. E. Sarachik, "Stability of circuits with randomly time varying parameters," IRE Trans. on Circuit Theory, Vol. CT-6, special supplement, pp. 260-270, May 1959.

[10] R. S. Bucy, "Stability and positive supermartingales," J. of Differential Equations, Vol. 1, pp. 151-155, 1965.

[11] H. J. Kushner, Stochastic Stability and Control, New York: Academic Press, 1967.

[12] R. E. Kalman and J. E. Bertram, "A unified approach to the theory of sampling systems," J. of the Franklin Institute, Vol. 267, pp. 405-436, May 1959.

[13] F. Kozin, "A survey of stability of stochastic systems," Automatica, Vol. 5, pp. 95-112, Pergamon Press, 1969.

[14] _____, "Stability of stochastic systems," Paper 3A, Third Congress IFAC, London, 1966.

[15] _____, "On relations between mements properties and almost sure Lyapunov stability for linear stochastic systems," J. Math Anal. Appl., Vol. 10, pp. 342-353, 1965.

[16] J. B. Pearson and J. E. Gibson, "On the asymptotic stability of a class of saturating sampled-data systems," Trans. Am. Inst. Elect. Engr., Vol. AIEE-84, (Applications and Industry), pp. 81-86, February 1964.

[17] E. I. Jury and B. W. Lee, "On the stability of a certain class of nonlinear sampled-data systems," IEEE Trans. on Automatic Control, Vol. AC-9, No. 1, January 1964.

[18] F. R. Gantmacher, The Theory of Matrices, New York: Chelsea Publications, 1959, translated from Russian by K. Hirsch.

[19] M. Loève, Probability Theory, Princeton, New Jersey: Van Nostrand, 1963, third edition.

[20] R. G. Agniel, "Stability of randomly sampled systems," Ph.D. Dissertation, Univ. of California, Berkeley, September 1969.

ACKNOWLEDGEMENT

†Research sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force under AFOSR Grant AF-AFOSR-68-1463 and the National Science Foundation under Grant GK-2277.

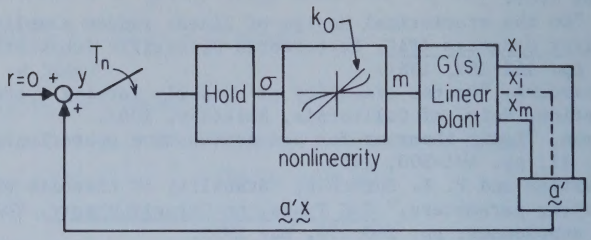


Figure 1: Nonlinear randomly sampled systems

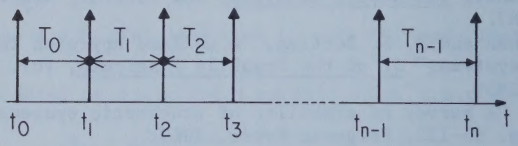


Figure 2: Sampling process

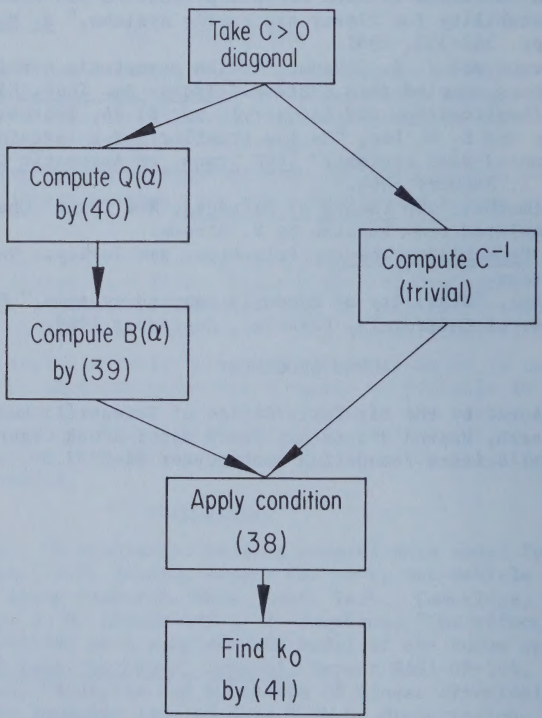


Figure 3: Flow chart for the computation of the largest k_0.

THE ABSOLUTE STABILITY OF NONLINEAR SYSTEMS INVOLVING A TIME-VARYING GAIN

T.L. CHANG and M.S. DAVIES
The University of British Columbia
Vancouver 8, B.C., Canada

ABSTRACT

The absolute stability of a class of nonlinear time-varying systems may be established by means of a frequency domain criterion. The criterion is a modified form of the familiar Popov condition.

INTRODUCTION

This paper treats the absolute stability of a class of nonlinear time-varying systems. The system is made up of three parts: a linear, time-invariant element, a sector nonlinearity, and a linear, time-varying gain. Such a system, in the absence of the time-varying part, falls into the class to which the well-known Popov stability criterion may be applied. Previous work [1], [2], has demonstrated that an augmented Popov-like criterion may be applied to time-varying systems and this paper also presents such a criterion. In some cases, as shown by example, the present result may usefully be applied to obtain sufficient conditions for stability. Some details of the proof have been omitted due to space limitations, a full version appears in [5].

THE STABILITY CONDITION

The system under consideration is shown in Figure 1. The input and zero-state output of the linear part are related by

$$\sigma_e(t) = \int_0^t g(t-\tau)e(\tau)\,d\tau .$$

The input $\eta(t)$ represents the zero-input response of G, together with any input function satisfying the conditions imposed below.

The input, $\sigma(t)$, and output, $\phi[\sigma(t)]$, of the memoryless nonlinearity are related in the following manner

$$\begin{aligned} &0 < \sigma\phi < \sigma^2 \quad \text{for } \sigma \neq 0,\ \phi(0) = 0 \\ &0 < \frac{\phi(\sigma_1) - \phi(\sigma_2)}{\sigma_1 - \sigma_2} \leqslant 1 \text{ for } \sigma_1 \neq \sigma_2 . \end{aligned} \tag{1}$$

The block k(t) represents a linear time-varying gain

$$c(t) = k(t)\,\phi[\sigma(t)].$$

The instantaneous value of this gain is constrained so that

(i) $K_1 < k(t) < K_2$ where $K_2 > K_1 > 0$

(ii) $b \leqslant \dot{k}(t) < ak(t)$ where $a > 0$ and b is finite.

It is assumed that the linear part is stable. More specifically:

(i) $g(t) \in L_2(0,\infty)$, $g(t) \in L_1(0,\infty)$

(ii) $\eta(t) \in L_1(0,\infty)$

(iii) η is differentiable and $\dot{\eta} \in L_1(0,\infty)$.

The above condition ensures that g is bounded on $(0,\infty)$ and that $g(t) \to 0$ as $t \to \infty$. The same is true of η.
Denote

$$\eta_M = \sup_{t \geq 0} |\eta(t)|, \qquad g_M = \sup_{t \geq 0} |g(t)|.$$

The Fourier transforms of $g(t)$, $e(t)$ etc. are denoted $G(j\omega)$, $E(j\omega)$ etc.
The notation $\|.\|$ denotes L_1 norms, thus

$$\|\eta\| = \int_0^\infty |\eta(t)| \, dt$$

The main result of this paper is the following theorem.

Theorem:

Consider the figure shown in Fig. 1, to which the assumptions given above apply. Let $y(t)$ be any real valued function such that

(i) $y(t) = 0$, $t < 0$

(ii) $y(t) \leq 0$ for $t \geq 0$

(iii) $\|y\| < K_1/K_2$

and let α be any positive number, then if

$$\text{Re}\left\{(1 + \alpha j\omega + Y(j\omega))(G(j\omega) + \frac{1}{K_2}) + a\alpha G(j\omega)\right\} - \|y\| \left(\frac{1}{K_1} - \frac{1}{K_2}\right) \geq 0$$

for all ω

then:

(i) $\sup_{t \geq 0} |\sigma(t)| < \infty$

(ii) $\sigma(t) \to 0$ as $t \to \infty$

(iii) as $\|\eta\| + \|\dot{\eta}\| \to 0$, the corresponding σ has the property that $\sup_{t \geq 0} |\sigma(t)| \to 0$

Proof

Define $$\sigma_T(t) = \begin{cases} \sigma(t) & t \leq T \\ 0 & t > T \end{cases}$$

$$c_T(t) = k(t)\,\phi[\,\sigma_T(t)]$$

$$\sigma_{eT}(t) = -\int_0^t g(t-\tau)c_T(\tau)\,d\tau \qquad (2)$$

Thus

$$\sigma_{eT}(t) = \sigma_e(t) \quad \text{for} \quad t \leqslant T \text{ and}$$

$$\sigma_{eT}(t) \;\varepsilon\; L_1\,(T,\infty)$$

The notation x * y denotes convolution, thus

$$x*y = \int_0^t x(\tau)y(t-\tau)\,d\tau$$

Define

$$\sigma_m = \sigma + \sigma*y,\quad c_m = c + c*y$$

Then

$$\int_0^T \left(\sigma_m(t) - \frac{c_m(t)}{K_2}\right) c(t)\,dt = \int_0^T \left(\sigma(t) - \frac{c(t)}{K_2}\right) c(t)\,dt$$

$$+\int_0^T [y * (\sigma - \frac{c}{K_2})]\,(t)\,c(t)\,dt \qquad (3)$$

Define

$$R(\tau) = \int_0^\infty [\,\sigma_T(t-\tau) - \frac{1}{K_2}\,c_T(t-\tau)]\;c_T(t)\,dt \qquad (4)$$

Now $R(\tau) = R_1(\tau) + R_2(\tau)$ where

$$R_1(\tau) = \int_0^\infty [\,\sigma_T(t-\tau) - \phi[\,\sigma_T(t-\tau)]]\;c_T(t)\,dt \qquad (5)$$

and

$$R_2(\tau) = \int_0^\infty [1 - \frac{k(t-\tau)]}{K_2}\;\phi[\,\sigma_T(t-\tau)]\;c_T(t)\,dt \qquad (6)$$

It may be shown that

$$R_1(\tau) \leqslant \int_0^\infty \frac{K_2}{K_1}\left\{\sigma_T(t) - \frac{c_T(t)}{K_2}\right\} \quad c_T(t)\,dt \qquad (7)$$

and also

$$R_2(\tau) \leqslant \int_0^\infty (\frac{1}{K_1} - \frac{1}{K_2})c_T(t)^2\,dt \qquad (8)$$

Thus

$$R(\tau) \leqslant \int_0^\infty (\frac{K_2}{K_1}\;\sigma_T(t) - \frac{c_T(t)}{K_2})c_T(t)\,dt \qquad (9)$$

Consider the last term of (3)

$$\int_0^T [y^*(\sigma - \frac{c}{K_2})](t)c(t)dt = \int_0^T \int_0^t y(\tau)[\sigma(t-\tau) - \frac{c(t-\tau)}{K_2}]c(t)d\tau dt$$

$$= \int_0^\infty \int_0^\infty y(\tau)(\sigma_T(t-\tau) - \frac{c_T(t-\tau)}{K_2})c_T(t)d\tau dt$$

$$= \int_0^\infty y(\tau)R(\tau)d\tau \qquad (10)$$

But, employing (9), we obtain

$$\int_0^T [y^*(\sigma - \frac{c}{K_2})](t)c(t)dt \geqslant \int_0^\infty y(\tau)d\tau \int_0^\infty (\frac{K_2}{K_1}\sigma_T(t) - \frac{c_T(t)}{K_2})c_T(t)dt \qquad (11)$$

$$\geqslant -\|y\| \int_0^\infty (\frac{K_2}{K_1}\sigma_T(t) - \frac{c_T(t)}{K_2})c_T(t)dt \qquad (12)$$

Substituting (12) in (3) yields

$$\int_0^T (\sigma_m(t) - \frac{c_m(t)}{K_2})c(t)dt + (\frac{1}{K_1} - \frac{1}{K_2})\|y\| \int_0^T c_T(t)^2 dt$$

$$\geqslant (1 - \frac{K_2}{K_1}\|y\|) \int_0^T (\sigma_T(t) - \frac{c_T(t)}{K_2})c_T(t)dt \qquad (13)$$

and, by the assumptions of the theorem, the right-hand side of (13) is non-negative.

Hence

$$\int_0^T (\sigma_m(t) - \frac{c_m(t)}{K_2})c(t)dt + (\frac{1}{K_1} - \frac{1}{K_2})\|y\| \int_0^T c_T(t)^2 dt \geqslant 0 \qquad (14)$$

Consider the integral

$$I = \int_0^T [\sigma_{em}(t) + \alpha\dot{\sigma}_e(t) - \frac{c_m(t)}{K_2} + a\alpha\sigma_e(t)$$

$$+ (\frac{1}{K_1} - \frac{1}{K_2})\|y\| c(t)]c(t)dt = I_1 + I_2 \qquad (15)$$

where

$$I_1 = \int_0^T [\sigma_{em}(t) - \frac{c_m(t)}{K_2} + (\frac{1}{K_1} - \frac{1}{K_2})\|y\| c(t)]c(t)dt \qquad (16)$$

and

$$I_2 = \int_o^T \alpha(\dot{\sigma}_e(t) + a\,\sigma_e(t))\,c(t)dt \tag{17}$$

But

$$I_1 = \int_o^T [\,\sigma_m(t) - \frac{c_m(t)}{K_2} + (\frac{1}{K_1} - \frac{1}{K_2})\,\|y\|\,c(t)]c(t)dt - \int_o^T \eta_m(t)c(t)dt \tag{18}$$

giving, due to (14)

$$I_1 \geq -\int_o^T \eta_m(t)c(t)dt \tag{19}$$

In a similar manner,

$$I_2 = \int_o^T \alpha(\dot{\sigma}(t) - a\,\sigma(t))c(t)dt - \int_o^T \alpha(\dot{\eta}(t) + a\eta(t))c(t)dt \tag{20}$$

It may be shown that

$$I_2 \geq -\alpha k(0)\int_o^{\sigma(o)} \phi(t)d(\sigma(t)) - \int_o^T \alpha(\dot{\eta}(t) + a\,\eta(t))c(t)dt \tag{21}$$

From (15),

$$I \int_o^\infty [\,\sigma_{eTm}(t) + \alpha\,\sigma_{eT}(t) - \frac{c_{Tm}(t)}{K_2} + a\alpha\,\sigma_{eT}(t) + (\frac{1}{K_1} - \frac{1}{K_2})\|y\|c_T(t)]c_T(t)dt \tag{22}$$

and, by the conditions of the theorem, since

$$I = \frac{-1}{2\pi}\int_{-\infty}^{\infty} \mathrm{Re}\{[1 + \alpha j\omega + y(j\omega)][G(j\omega) + \frac{1}{K_2}] + a\alpha G(j\omega) - (\frac{1}{K_1} - \frac{1}{K_2})\|y\|\}\,C_T(j\omega)C_T(-j\omega)d\omega$$

it follows that $I \leq 0$ (23)

Since $I = I_1 + I_2$, using (21) and (19) substituted in (23)

$$\alpha[k(T)\Phi(T) - k(o)\Phi(o)] \leq \int_o^T (a\alpha\eta(t) + \eta_m(t) + \alpha\,\dot{\eta}(t))c(t)dt. \tag{24}$$

Defining $c_M = \sup_{o<t<T} |c(t)|$ and invoking the conditions already imposed upon η, $\dot{\eta}$, and y, the right-hand side of (24) must be less than the quantity

$$c_M[(1 + \|y\| + a\alpha)\|\eta\| + \alpha\|\dot{\eta}\|] = M\,c_M \tag{25}$$

Further, considering the first term of the left-hand side of (24), since ϕ is monotonic

$$\Phi(t) \geq 1/2\ \{\phi[\sigma(t)]\}^2 \tag{26}$$

Using (25) and (26) in (24) yields:

$$\frac{\alpha}{2} k(T)[\phi(\sigma(T))]^2 \leqslant \alpha k(o)\Phi(o) + M c_M \tag{27}$$

From (27), since $\frac{k(T)}{K_2} \leqslant 1$

$$\frac{\alpha}{2K_2} c(T)^2 \leqslant \alpha k(o)\Phi(o) + M c_M \tag{28}$$

The inequality (28) holds for any $T \geqslant 0$ and implies

$$\sup_{t \geqslant 0} |c(t)| \leqslant \frac{K_2 M}{\alpha} + \left[\left(\frac{K_2 M}{\alpha^2}\right)^2 + 2K_2 k(o)\Phi(o)\right]^{1/2} \tag{29}$$

Further, this bound on c(t) tends to zero with $\|\eta\| + \|\dot{\eta}\|$.

It is now shown that $|c(t)| \to 0$ as $t \to \infty$. Once again using (23), and substituting in (13) and (21).

$$(1 - \frac{K_2}{K_1}\|y\|)\int_o^T (\sigma(t) - \frac{c(t)}{K_2})c(t)dt - \int_o^T \eta_m(t)c(t)dt - \alpha k(o)\Phi(o)$$

$$- \alpha \int_o^T (\dot{\eta}(t) + a\,\eta(t))c(t)dt \leqslant 0 \tag{30}$$

Thus

$$(1 - \frac{K_2}{K_1}\|y\|)\int_o^T (\sigma(t) - \frac{c(t)}{K_2})c(t)dt \leqslant \alpha k(o)\Phi(o) + M c_M \tag{31}$$

From which it follows that

$$\int_o^\infty (\sigma(t) - \frac{c(t)}{K_2})c(t)dt < (\alpha k(o)\Phi(o) + M c_M)/(1 - \frac{K_2}{K_1}\|y\|) \tag{32}$$

and this bound on the integral tends to zero with $\|\eta\| + \|\dot{\eta}\|$.

However, noting that the bound already placed on c, together with the conditions required of $\dot{g}(t)$ require $\dot{\sigma}_e(t)$ to be bounded, it may be shown that the integral of (32) is infinite unless $\sigma(t) \to 0$ as $t \to \infty$, contradicting (32).

Therefore, it is concluded that $\sigma(t) \to 0$ as $t \to \infty$. In addition, since $c \to 0$ only if $\sigma \to 0$, the previous bound on c implies that $|\sigma(t)| \to 0$ as $\|\eta\| + \|\dot{\eta}\| \to 0$.

Example

Referring to Fig. 1, assuming the linear part has a transfer function

$$G(s) = \frac{K}{(s+1)(s+2)}.$$

The nonlinear part satisfies (1), and the time varying gain is such that

(i) $1 < k(t) < 1.2$

(ii) $+b \leqslant k(t) \leqslant a\ k(t)$

where $+b > -\infty$, $a < \infty$, a and b are very large arbitrary numbers.

Taking $\alpha = 0$. Suppose $y(t)$ is an exponential function,

$$y(t) = -\frac{1}{\gamma} e^{-\frac{\beta t}{\gamma}} \quad \text{with } t \geqslant 0,\ \gamma > 0, \text{ and } \beta > \frac{K_2}{K_1} = 1.2,$$

$$\| y(t) \| = \int_0^\infty \left| \frac{-1}{\gamma} e^{-\frac{\beta t}{\gamma}} \right| dt = + \frac{1}{\beta} < \frac{K_1}{K_2}$$

Taking Laplace transforms

$$1 + y(s) = 1 - \frac{1}{\gamma} \frac{1}{s + \frac{\beta}{\gamma}} = 1 - \frac{1}{\beta + \gamma s}$$

Now, the stability condition is that

$$\text{Re} \left\{ \left(1 - \frac{1}{\beta + \gamma s}\right) \left[G(s) + \frac{1}{1.2} \right] - \frac{1}{\beta} \frac{0.2}{1.2} \right\} > 0 \qquad \text{for all } \omega.$$

The locus of $G'(j\omega) = \left(1 - \frac{1}{\beta + \gamma j\omega}\right) [G(j\omega) + 0.833]$

must lie on the right side of the vertical line passing through the point $(0.1667/\beta, 0)$. The first part of $G'(j\omega)$ can be considered a compensator.

Choosing $\beta = 2$, $\gamma = 0.5$ is reasonable, but is not best. It can be shown that the maximum K is 23.75. From Fig. 2 $G'(j\omega)$ locus lies in the right side of the vertical line passing point $(0.08333, 0)$ and the system satisfies the stability condition.

CONCLUSION

The main difficulty in testing the stability criterion lies in choosing a function $y(\cdot)$ and parameter α. As shown in an example, limited success may be obtained with simple forms of $y(\cdot)$.

It is evident that if the parameter $\dot{k}$ is taken as zero; the criterion reduces to that of Baker and Desoer for time-invariant systems.

The generalization to the multivariable case may be made in a straightforward manner.

This research was supported in part by Research Grant A-4148 of the National Research Council of Canada.

REFERENCES

1. Zames, G.: On the input-output stability of time varying nonlinear feedback systems, Part I, II, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228-238, 465-476.

2. Zames, G. and P.L. Falb: Stability Conditions for systems with monotone

and slope-restricted nonlinearities, SIAM J. Control, Vol. 6, No. 1 (1968), pp. 89-108.

3. Baker, R.A. and Desoer, C.A.: Asymptotic stability in the large of a class of single-loop feedback systems, SIAM J. Control, Vol. 6, No. 1, (1968), pp. 1-8.

4. Bergen, A.R. and Rault, A.J.: Absolute input-output stability of feedback systems with single time varying gain, Journal of the Franklin Institutes, Vol. 286, No. 4, (1968), pp. 280-294.

5. Chang, T.L.: The Absolute Stability of Nonlinear Systems, M.A.Sc. Thesis, Department of Electrical Engineering, University of British Columbia, Vancouver.

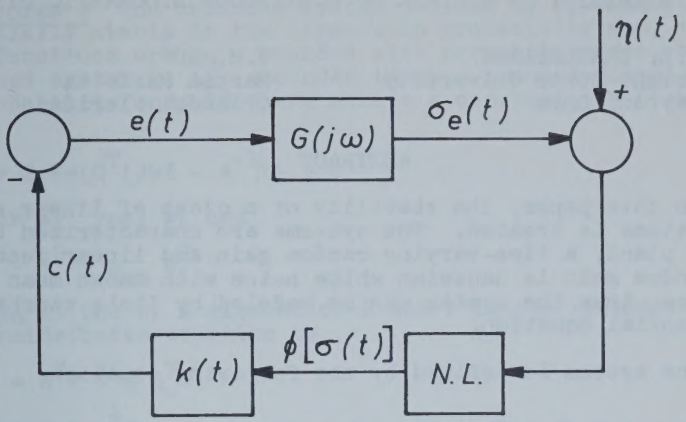


FIGURE 1

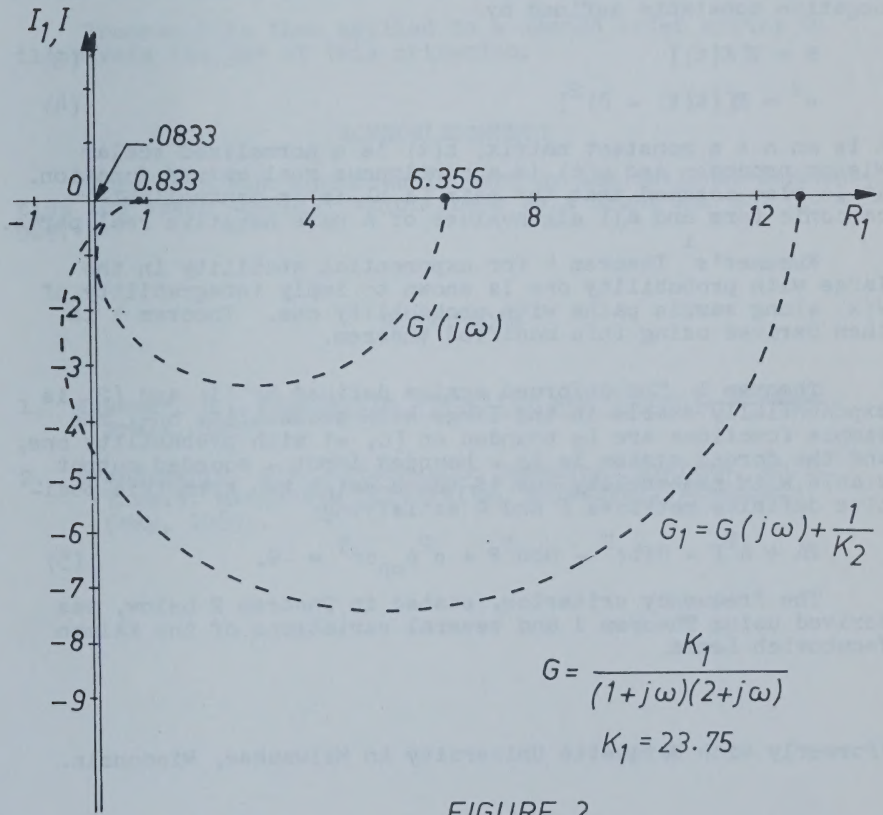


FIGURE 2

ON THE STABILITY OF A CLASS OF CONTINUOUS STOCHASTIC SYSTEMS

J.E. BRANDEBERRY
Wright State University
Dayton, Ohio

S.H. WU*
Martin Marietta
Orlando, Florida

ABSTRACT

In this paper, the stability of a class of linear stochastic systems is treated. The systems are characterized by a linear plant, a time-varying random gain and linear feedback. The random gain is Gaussian white noise with known mean and variance; thus the system can be modeled by Ito's stochastic differential equation.

The system is defined by the following equation:

$$dx = A\,x\,dt + \beta b\,e\,dt + \sigma b\,e\,dz \tag{1}$$

$$e(t) = u(t) - c^T x \tag{2}$$

where x, b, and c are n-dimensional vectors; β and σ are non-negative constants defined by

$$\beta = E[K(t)] \tag{3}$$

$$\sigma^2 = E[(K(t) - \beta)^2] \tag{4}$$

A is an $n \times n$ constant matrix; z(t) is a normalized scalar Wiener process; and u(t) is a continuous real valued function. It is also assumed that the pair [A,b] is in controllable canconic form and all eigenvalues of A have negative real parts.

Kushner's[1] Theorem 4 for exponential stability in the large with probability one is shown to imply integrability of V(x) along sample paths with probability one. Theorem 1 is then derived using this modified theorem.

Theorem 1 The unforced system defined by (1) and (2) is exponentially stable in the large with probability one, the sample functions are L_2 bounded on $[0, \infty)$ with probability one, and the forced system is L_2 - bounded input - bounded output stable with probability one if there exist two symmetric positive definite matrices P and Q satisfying:

$$PA + A^T P - \beta P b c^T - \beta c b^T P + \sigma^2 p_{nn} c c^T = -Q. \tag{5}$$

The frequency criterion, stated in Theorem 2 below, was derived using Theorem 1 and several variations of the Kalman Yacubovich Lemma.

*Formerly with Marquette University in Milwaukee, Wisconsin.

Theorem 2 The unforced system defined by (1) and (2) is exponentially stable in the large with probability one, the sample functions are L_2 - bounded with probability one, and the forced system is L_2 - bounded input - bounded output stable with probability one; if there exists a constant $\gamma \geq 0$ such that

$$\gamma + 2\,\mathrm{Re}(C^T(jwI - A^{-1}b) > 0 \tag{6}$$

holds for all real w, and

$$2\beta > \beta^2 + p_{nn}\,\sigma^2 \tag{7}$$

where p_{nn} is the n, n element of P and P is the symmetric positive semidefinite solution to

$$PA + A^TP = q\,q^T \tag{8}$$

$$Pb = c + \gamma^{\frac{1}{2}}q. \tag{9}$$

It is also possible to find p_{nn} by spectral factoring as in the proof of the Kalman Yacubovich Lemma. Comment: For detail proofs of Theorems 1 and 2 see Reference 2.

Theorem 2 is then applied to a second order system to illustrate the use of this criterion.

ACKNOWLEDGMENTS

This work was supported in part by the National Aeronautics and Space Administration under NASA Grant No. NGT 50-001-005.

REFERENCES

1. Kushner, H., Stochastic Stability and Control, Academic Press, 1967.

2. Brandeberry, J.E., Stochastic Dynamical Systems, Ph.D. Thesis, Marquette University, Milwaukee, Wisconsin (May, 1969).

SET STABILITY OF DISCRETE DYNAMICAL SYSTEMS

JAMES A. HEINEN
Dept. of Electrical Engineering
Marquette University
Milwaukee, Wisconsin

SHERMAN H. WU
Martin Marietta Corp.
Orlando, Florida

ABSTRACT

A new type of stability (called "set stability" and involving specific bounds on solutions) is introduced for discrete systems. Liapunov-like theorems yielding sufficient conditions for set stability and uniform set stability are presented.

INTRODUCTION

The discrete dynamical systems to be considered are first order n-dimensional vector difference equations of the form

$$x(j+1) = f(x(j),j) \qquad (D)$$

where $f:R^n \times J \to R^n$, R^n is n-dimensional Euclidean space, and $J = \{j_i, j_i+1, \cdots, j_f\}$ is a finite or infinite subset of the integers. Unlike the usual qualitative forms of stability, the type of stability to be considered here is quantitative in nature. In the following definitions S and S_o are time-varying subsets of R^n (i.e., $S, S_o: J \to P(R^n)$, the power set of R^n) with the exception that S_o in Definition 1 is constant.

STABILITY DEFINITIONS

Definition 1. A discrete system (D) is set stable with respect to S, S_o, and J (abbreviated $SS(S,S_o,J)$) if, whenever $x(j)$ is a solution of (D), $x(j_i) \varepsilon S_o$ implies $x(j) \varepsilon S(j)$ for all $j \varepsilon J$.

Definition 2. A discrete system (D) is uniformly set stable with respect to S, S_o, and J (abbreviated $USS(S,S_o,J)$) if, whenever $x(j)$ is a solution of (D) and $j_o \varepsilon J$, $x(j_o) \varepsilon S_o(j_o)$ implies $x(j) \varepsilon S(j)$ for all $j \varepsilon J$, $j \geq j_o$.

In the special case where S and S_o are chosen to be constant spheres in some norm topology and J is finite, $SS(S,S_o,J)$ reduces to Michel and Wu's [1] finite time stability for discrete systems.

STABILITY THEOREMS

The following Liapunov-like theorems yielding sufficient conditions for set stability and uniform set stability can be proven by contradiction. In these theorems $\Delta V(x,j) = V(f(x,j),j+1) - V(x,j)$.

Theorem 1. Suppose $S: J \to P(R^n)$ and $S_o \subset S(j_i)$. If there exist a function $V: R^n \times J \to R^1$, a real (finite) constant M, and a function $m: J \to R^1$ such that

(a) $\Delta V(x,j) \leq 0$ for all $x \varepsilon S(j)$ and all $j_i \leq j < j_f$,
(b) $m(j) \leq V(x,j)$ for all $x \varepsilon \tilde{S}(j)$ (the complement of $S(j)$) and all $j_i < j \leq j_f$,
(c) $V(x,j_i) \leq M$ for all $x \varepsilon S_o$,
(d) $M \leq m(j)$ for all $j_i < j \leq j_f$, and
(e) strict inequality holds in at least one of the four conditions (a) - (d),

then the discrete system (D) is $SS(S,S_o,J)$.

Theorem 2. Suppose $S, S_o : J \to P(R^n)$ are such that $S_o(j) \subset S(j)$ for all $j \in J$. If there exist a function $V: R^n \times J \to R^1$ and functions $m, M: J \to R^1$ such that

(a) $\Delta V(x,j) \leq 0$ for all $x \in S(j)$ for all $j_i \leq j < j_f$,
(b) $m(j) \leq V(x,j)$ for all $x \in S(j)$ for all $j_i < j \leq j_f$,
(c) $V(x,j) \leq M(j)$ for all $x \in S_o(j)$ for all $j_i \leq j < j_f$,
(d) $M(j_1) \leq m(j_2)$ for all $j_i \leq j_1 < j_2 \leq j_f$, and
(e) strict inequality holds in at least one of the four conditions (a) - (d),

then the discrete system (D) is $USS(S, S_o, J)$.

CONCLUDING REMARKS

The definitions and theorems for set stability are quite general. No assumptions regarding uniqueness of solutions of (D), the classical stability of (D), the continuity of V, or the boundedness of S and S_o are made. The theorems can be applied to linear, nonlinear, and time-varying systems defined on both finite and infinite time intervals.

REFERENCE

[1] Michel, A. N., and S. H. Wu, "Stability of discrete systems over a finite interval of time," Int. J. Control, Vol. 9, No. 6, pp. 679-693, June 1969.

ON THE STABILITY OF LURIE SYSTEMS WITH STOCHASTIC INPUT

S.PURKAYASTHA
Technical Officer
Civil Aviation
New Delhi

A.K.MAHALANABIS
Assistant Professor
Inidian Institute of Technology
New Delhi

ABSTRACT

The problem of stability of a Lurie type nonlinear feedback system having a Gaussian random input is considered. The method of V.M.Popov is modified in order to derive a sufficient condition for the system to be asymptotically stable.A numerical example is included in order to illustrate the applicability of the result.

INTRODUCTION

Various aspects of the problem of stability of Lurie type nonlinear feedback systems,with or without external input, have been subjected to intensive studies over the past several years.However,most of these have been concerned with deterministic situations.Three of the possible stochastic versions of these topics have been considered by Morozan[1],Wonham[2], and Naumov[3].Morozan has studied autonomous systems having random parameters and has been able to extend the method of V.M.Popov for deriving sufficient conditions for these systems to be asymptotically stable in the mean.Wonham,on the other hand,has considered the effect of a white noise input on the 'positivity',an weak counterpart of the Lagrange stability of the determistic systems.He has made use of the Ito calculus along with a stochastic version of the Liapunov's second method. The result in[2]indicates that the Popov-Kalman inequality for the absolute stability of the autonomous Lurie system also provides a sufficient condition for the system with white noise input to be positive.

Wonham's result,though useful for systems proved asymptotically stable in the autonomous state,does not help to explain the well known dependence of a nonlinear system's stability on external input.The result in[3]is an improvement in this respect.Assuming the input to be Gaussian with a finite mean value,the author has derived a nonlinear integral equation governing the evolution of the mean value of the output.This has been shown to correspond to a non-autonomous deterministic

system and a stability condition has been obtained from the results of Tsypkin and Naumov[4].

In the present paper the problem of Naumov is re-examined. For simplicity, the input has been assumed to be a zero mean Gaussian signal with its lowest spectral component located well beyond the system passband. This ensures that the output of the system is deterministic and that there is no feedback at the input signal frequencies. However, this is not an essential condition for the following analysis to be useful. If an approximation is made, as in [3], by assuming the output distribution to be Gaussian (with unknown mean and variance) the same stability condition can be derived by a slight modification of the procedure given here even if the input has low frequency components.

PRELIMINARIES

The system under consideration is shown in Fig.1. The input $r(t)$ (the dependence of the random signals on the sample variable will be omitted for notational convenience) is a high frequency Gaussian signal with mean zero and variance R. The function $f(e)$ stands for the controller characteristic which is assumed to be of the zero memory type, continuous or discontinuous, and satisfies the conditions: $f(0) = 0$, $0 \leq f(e) \leq ke$, $e \neq 0$ with $k \in (0, \infty)$. The linear sub-system is represented by its transfer function $G(s)$ and is assumed to be stable, completely controllable and completely observable. However, as in the case of autonomous systems, the analysis can be simply extended to include systems with $G(s)$ semistable.

The output $y(t)$ can be expressed by the nonlinear integral equation

$$(1) \qquad y(t) = y_o(t) - \int_0^t g(t-z)f(z)dz$$

where $y_o(t)$ is the response to initial conditions, $g(t)$ is the impulse response function of the linear sub-system and $f(t)$ is written for $f[e(t)]$. The error signal $e(t)$ is given by

$$(2) \qquad e(t) = r(t) - y(t)$$

Obviously $e(t)$ is also Gaussian with a sample mean $-y(t)$ and variance R.

We shall use the subscript m to indicate the sample

averages of the random processes.Thus the following new variables are introduced :

$$(3) \qquad e_m(t) = E\, e(t) = -y(t)$$

$$(4) \qquad f_m(t) = E\, f(t)$$

where E . denotes the expectation operation with respect to the distribution of e(t).It is assumed that the characteristic $f_m(e_m)$ satisfy the conditions : $f_m(0) = 0$, $0 \leq f_m(e_m) \leq k_m e_m$, $e_m \neq 0$.

It is now required to introduce the definition of the stability investigated in the paper.

Definition.

The system will be said to be globally asymptotically stable in the mean (GASIM) if for a given R and for all initial disturbances the output y (t) tends to zero as t tends to ∞.

Remark 1.

If the input r(t) has low frequency components to cause the output y(t) to be random also,the right hand side of (3) need be replaced by $-y_m(t)$,the expected value of the output. Then the definition above has to be modified by replacing y(t) by $y_m(t)$.

Problem statement.

The problem considered in the paper is to derive a frequency domain sufficient condition for the system in Fig.1 to be GASIM for a specified value of R.

THE MAIN RESULT

The main result established in the paper can be stated in the form of the following theorem:

Theorem.

Consider the system (1) with G(s) stable ,completely controllable and completely observable ,the input and the controller satisfying the conditions stated in the preceding section.A sufficient condition for the system to be GASIM is that there exists a real constant q such that the following inequality is satisfied for all real frequencies $\omega \in [0, \infty)$:

$$(5) \qquad \mathrm{Re}(1 + j\omega q)G(j\omega) + 1/k_m \geq 0$$

Remark 2.

Inequality (5) has the form of the Popov inequality for the absolute stability of the system with $r(t) = 0$. The point of distinction is the replacement of the sector gain k of the nonlinearity $f(e)$ by the gain k_m of the mean characteristic $f_m(e_m)$. In general, $k_m \leq k$ so that inequality (5) can be satisfied for a given system even if the Popov condition is not met in the autonomous state.

Remark 3.

The inequality derived in [3] has the form

$$\text{Re } G(j\omega) + 1/k_m \geq 0 \tag{6}$$

As is well known, inequality (5) is, in general, less restrictive than inequality (6). This indicates the improvement achieved in the present analysis.

Remark 4.

As in the case of autonomous systems, it is possible to modify the proof of the main result to follow so as to derive a number of useful corollaries. For example, if $G(s)$ is semistable, it can be shown that inequality (5) along with the condition of limit stability ensures the system to be GASIM. Also, stability conditions can be derived under the assumption of known constraints on the slope of the nonlinearity, and so on.

PROOF OF THE MAIN RESULT

Define the following two truncated functions for an arbitrary $T \in [0, \infty)$:

$$f_{mT} = f_m(t), \quad 0 \leq t \leq T \tag{7}$$
$$= 0, \quad t > T$$

$$v_T(t) = y_0(t) - y(t), \quad 0 \leq t \leq T \tag{8}$$
$$= -\int_0^T g(t-z)f(z)dz, \quad t > T$$

Evidently, the function $v_T(t)$ can be expressed, for all t, as

$$v_T(t) = -\int_0^t g(t-z)f_T(z)dz \tag{9}$$

Taking expectation of both sides of (9), it can be seen that the function $v_T(t)$ can also be written as

$$(10) \qquad v_T(t) = -\int_0^t g(t-z) f_{mT}(z)\,dz$$

If we take Fourier transform of both sides of (10) we get the relation

$$(11) \qquad \mathcal{F}\{v_T(t)\} = -G(j\omega)F_m(j\omega)$$

where

$$(12) \qquad F_m(j\omega) = \mathcal{F}\{f_{mT}(t)\}$$

Also, from the well known property of Fourier transforms

$$(12) \qquad \mathcal{F}\{\dot{v}_T(t)\} = -j\omega\, G(j\omega)F_m(j\omega)$$

Now, consider the function $\rho(T)$ defined by

$$(13) \qquad \rho(T) = \int_0^{\infty} \left[v_T - f_{mT}/k_m + q\dot{v}_T\right] f_{mT}\,dt$$

Applying Parseval's theorem the right hand side of (13) can be expressed in terms of the transformed quantities to get

$$(14) \qquad \rho(T) = -(1/2\pi)\int_{-\infty}^{\infty} \left[\mathrm{Re}\,(1+j\omega q)G(j\omega) + 1/k_m\right] |F_m(j\omega)|^2\,d\omega$$

Assuming the validity of the theorem it can be seen from (14) that inequality (5) implies that

$$(15) \qquad \rho(T) \leq 0$$

Substituting for $v_T(t)$ from (8) in (13) and rearranging the terms it can be seen that (15) implies

$$(16) \qquad \int_0^T (v_T - f_{mT}/k_m) f_{mT}\,dt + q\int_{v_T(0)}^{v_T(T)} f_{mT}\,dv_T \leq 0$$

By the assumption on the characteristic $f_m(e)$

$$f_m(t) \leq -k_m y(t)$$

Thus inequality (16) will be satisfied if the following is satisfied:

$$(17) \qquad \int_0^T (-y - f_m/k_m)(-k_m y)\,dt + qk_m \int_{y(0)}^{y(T)} y\,dy \leq -\int_0^T (y_o + q\dot{y}_o) f_m\,dt$$

Apparently inequality (17) has the form

$$(18) \qquad c_1\int_0^T y^2\,dt + c_2\int_{y(0)}^{y(T)} y\,dy \leq -c_3\int_0^T (y_o + q\dot{y}_o) f_m\,dt$$

where the c_i's are all positive constants.

Inequality (18) is exactly the same as the one involved in the proof of Popov's theorem (see for example[5]).This proves the main result.

AN EXAMPLE

A numerical example is presented here.The results of the present analysis are compared with those obtained from the method of[3]and of simulation studies.

The system corresponds to a linear sub-system transfer function

$$G(s) = 4/s(s+1)^2$$

and a controller given as

$$f(e) = \text{Sgn } e$$

The signal r(t) is a zero mean random signal with spectral components in the range 100 c/s - 100 Kc/s.
It is seen that the autonomous version of the system fails to satisfy the Popov condition and sustains limit cycle oscillations.The system has been simulated on an analog computer and has been found to be stabilized(i.e.,there is no limit cycle oscillations) if the input RMS value is greater than 1.58.

To find the values of$\sqrt{R}$ from inequalities (5) and(6),it is required first to determine the characteristic $f_m(e)$.This is given by

$$f_m = 2\emptyset(y/\sqrt{R}) - 1 \quad , \emptyset \text{ is the probability integral}$$

Plots of this function for some representative values of R are shown in Fig.2.Once the value of $\sqrt{R}$ is chosen it is possible to pick up the relavent f_m characteristic and determine the corresponding sector.This gives the gain k_m which is needed in (5) and (6).

For the example system,the required value of$\sqrt{R}$ is found to be equal to 2 from (5) and equal to 8 from (6).

CONCLUSIONS

It is shown that the method of Popov can be modified to derive a meaningful frequency domain inequality for the mean asymptotic stability of a Lurie system with random input.

REFERENCES

1.T.Morozan,"The method of V.M.Popov for control systems with random parameters",J.Math.Anal.App.,16,pp.201-215,Nov 1966.

2.W.M.Wonham,"Liapunov criteria for weak stochastic stability", J.Diff.Equns.,2,pp.195-207,April 1966.

3.B.N.Naumov,"Frequency method for investigation of absolute stability and quality of nonlinear systems and systems with time variable parameters affected by given or random disturbances",paper no.23A,IFAC Congress,London,1966.

4.B.N.Naumov and Ya.Z.Tsypkin,"A frequency criterion for absolute process stability in nonlinear automatic control systems", Autom.Rem.Control,25,pp.765-778,June 1964.

5.M.A.Aizerman and F.R.Gantmakher,"Absolute stability of regulator systems",Holden Day Inc.,SanFrancisco,1964.

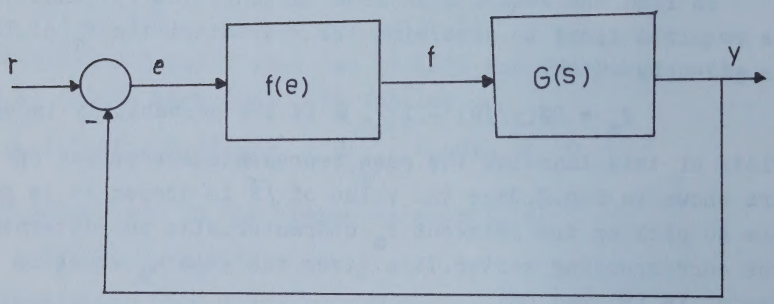


Fig. 1

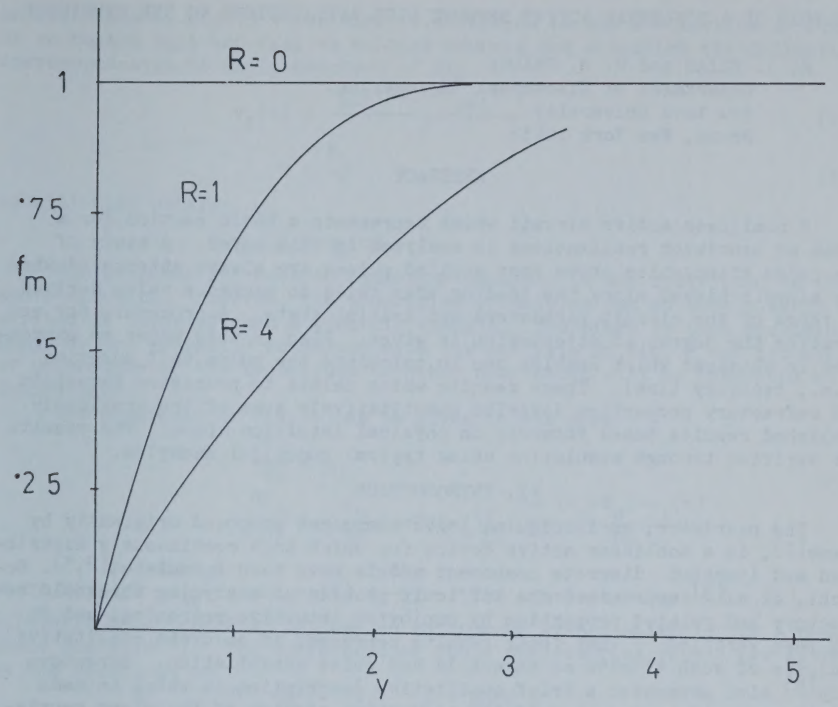


Fig. 2

ANALYSIS OF A NONLINEAR ACTIVE NETWORK WITH APPLICATIONS TO THE NEURISTOR

N. J. ELIAS and M. S. GHAUSI
Department of Electrical Engineering
New York University
Bronx, New York 10453

ABSTRACT

A nonlinear active circuit which represents a basic section for a class of neuristor realizations is analyzed in this paper. A study of the pulse attenuation shows that applied pulses are always attenuated when the slope achieved along the leading edge fails to exceed a value derived in terms of the circuit parameters and initial state. A procedure for estimating the degree of attenuation is given. Also in this paper an expression is obtained which enables one to calculate the pulse tail duration (i.e., recovery time). These results which relate to neuristor threshold and refractory properties describe quantitatively some of the previously published results based formerly on physical intuition alone. The results are verified through simulation using typical numerical examples.

I. INTRODUCTION

The neuristor, an intriguing logic component proposed originally by Crane[1], is a nonlinear active device for which both continuously distributed and iterated discrete component models have been formulated[2,3]. Noguchi, et al[3] approached the difficult problem of analyzing threshold refractory and related properties by employing intuitive reasoning, and it has been verified[4] that their results represent an accurate qualitative analysis of such effects as threshold and pulse annihilation. Berestovskiy[5] also presented a brief qualitative description in which he made reference to the properties of the elementary section of the given neuristor circuit. In this paper a quantitative analysis of such an elementary circuit (Fig. 1) will be performed.

Specific neuristor properties can be analyzed quantitatively by basing that analysis on the properties of the elementary circuit. Determination of threshold conditions previously based on establishing whether or not the nonlinear element is caused to exhibit negative conductance[3] can be more precisely formulated in terms of whether or not a pulse applied to the elementary section is attenuated despite such negative conductance. Rates of attenuation (per unit distance) exhibited in a refractory region of the neuristor can be quantitatively established by analyzing the degree of attenuation observed in the elementary section whereas previous reports[3] suggest the intuitive assumption of an exponential rate. The minimum necessary time interval between transmission of successive pulses in the neuristor about which little if anything has been reported can be determined using a formula for the duration of the tail of the response pulse in the elementary section. Thus we are interested in studying pulse attenuation and recovery time in this nonlinear active circuit.

II. THE CIRCUIT AND ITS EQUATIONS

The specific circuit to be analyzed in this paper is shown in Fig. 2. It can be regarded as the elementary section of the discrete component neuristor model (Fig. 2) proposed by Nagumo, Arimoto and Yoshizawa[2] who employed tunnel diodes to provide the necessary active behavior. The analysis to follow does not require reference to any specific nonlinear characteristic provided $i(v_o)$ is a physically realistic, single-valued function exhibiting negative resistance. For the sake of clarity, however, the reader may assume that a tunnel diode is implied.

Briefly the relation between the circuits of Figs. 1 and 2 follows up-

on reference to Fig. 3 which directs attention to the n^{th} section of Fig. 2. If we regard v_{n-1} and v_{n+1} as voltage sources and establish the following correspondences to the parameters of Fig. 1

$$v_i(t) \equiv \frac{v_{n-1}(t) + v_{n+1}(t)}{2} \tag{2.1A}$$

$$R \equiv \frac{R_1}{2} \tag{2.1B}$$

we will also obtain

$$v_o(t) \equiv v_n(t). \tag{2.1C}$$

The proposed approach to neuristor analysis is not restricted to the above circuits. However, the underlying principles can be most clearly described in terms of such a specific example. Therefore in this preliminary paper we are primarily concerned with the analysis of Fig. 2 and discussion of extensions of the techniques to other neuristors will be presented at a future date.

The equations of Fig. 1 can be given in the form

$$\begin{aligned} \frac{dv_o}{dt} &= y_o \\ \frac{dy_o}{dt} &= -\frac{y_o}{RC}[1+RG(v_o)] - \frac{\frac{R}{L}(v_o-E_o)-\dot{v}_i(t)}{RC} \end{aligned} \tag{2.2}$$

where $G(v_o) \equiv \frac{di(v_o)}{dt}$ and $\dot{v}_i(t) \equiv \frac{dv_i}{dt}$. For constant values of $\dot{v}_i(t)$, $\dot{v}_i(t)=\dot{V}_i$, Equations (2.2) become an autonomous system suitable for a phase plane analysis[6]. The corresponding phase trajectories $y_o(v_o)$ are given by the solutions of

$$\frac{dy_o}{dv_o} = -\frac{R(v_o-E_o)-L\dot{V}_i}{y_o\,RLC} - \frac{1+RG(v_o)}{RC} \tag{2.3}$$

which is obtained by dividing the first of Equations (2.2) into the second. Solving (2.3) for y_o we find that a given value of $\frac{dy_o}{dv_o} = m$ will be exhibited by a trajectory wherever it intersects the isocline

$$y_m(v_o) = -\frac{R(V_o-E_o) + L\dot{V}_i}{RLC\,m+L[1+RG(v_o)]} \tag{2.4}$$

In the analysis that follows we will make frequent use of these equations. Even when $\dot{v}_i(t)$ is not constant this procedure is of use since we can always regard $v_i(t)$ as the limiting case of a series of linear segments.

III. ATTENUATION OF PULSES

In Equations (2.2) the excitation $v_i(t)$ exerts its influence only through the appearance of its slope $\dot{v}_i(t)$. This indicates that in the neuristor of Fig. 2 threshold will depend on the trigger slope $\frac{dv_{in}}{dt}$ (and duration) and that pulse attenuation can be observed by comparison (on identical scales) of the pulse slopes exhibited at successive sections. For the circuit of Fig. 1 we will, therefore, establish the conditions under which $\frac{dv_o}{dt} < \frac{dv_i}{dt}$.

Assume that the circuit is initially unexcited and at equilibrium with $v_o(0)=E_o$ and $y_o(0)=0$. At $t=0$ a signal $v_i(t)$ is applied for which $v_i(t)$ is maintained at a constant level $\dot{V}_i$ for at least as long as it takes $y_o(t)$ to

reach a maximum value (more general excitations will be introduced subsequently). Let α be some prescribed (attenuation) factor. Our aim is to establish a quantity $\dot{V}_T$ in terms of the circuit parameters such that

$$\frac{dv_o}{dt} \leq \alpha \dot{V}_i \tag{3.1}$$

whenever $0 < \dot{V}_i \leq \dot{V}_T$.

To determine an expression for $\dot{V}_T$ we refer to the phase plane to seek conditions under which (3.1) applies. Fig. 4 depicts the technique to be used. Along with the trajectory, $y_o(v_o)$, we have sketched a "test curve"

$$y_o = \alpha \dot{V}_i f(v_o) \tag{3.2A}$$

where

$$f(v_o) = 1 - \frac{1-k}{(1-k)(v-E_o) + 1} \tag{3.2B}$$

with k a constant value between 0 and 1 (to be selected shortly) such that this test curve rises monotonically from a positive value $k\alpha\dot{V}_i$ at $v=E_o$ and approaches $\alpha\dot{V}_i$ from below. The trajectory of interest emanates from $v_o=E_o$, $y_o=0$ (below the test curve) and rises towards increasing y_o and v_o proceeding in a clockwise direction for increasing time. Inequality (3.1) will apply for all values of $\dot{V}_i$ such that this trajectory remains below the test curve. This will certainly be the case if no trajectory at its intersection with the test curve exhibits a slope, $\left.\frac{dy_o}{dv_o}\right|_{y_o=\alpha\dot{V}_i f(v_o)}$, in excess of the test curve slope, $\frac{d}{dv_o}(\alpha\dot{V}_i f(v_o))$, for then all trajectories intersecting the test curve must approach that curve from above.

Thus inequality (3.1) applies for all values of $\dot{V}_i$ such that on a pointwise basis

$$\frac{d}{dv_o}(\alpha\dot{V}_i f(v_o)) - \left.\frac{dy_o}{dv_o}\right|_{y_o=\alpha\dot{V}_i f(v_o)} > 0 \qquad v_o \geq E_o$$

With the second term determined from Equation (2.3) and $\frac{df(v_o)}{dv_o}$ denoted by $f'(v_o)$ this condition becomes

$$\alpha\dot{V}_i f'(v_o) + \frac{R(v_o-E_o)-L\dot{V}_i}{\alpha RLC\,\dot{V}_i f(v_o)} + \frac{1+RG(v)}{RC} > 0 \qquad v_o \geq E_o \tag{3.3A}$$

Note that for $v_o \to E_o$ and $\dot{V}_i \to 0$ (3.3A) yields

$$f(E_o) = k > \frac{1}{\alpha[1+R\,G(E_o)]} .$$

Aside from this restriction k can be selected for the optimum result (i.e., largest $\dot{V}_T$).

Our problem is now to find a $\dot{V}_T$ such that (3.3A) follows whenever $0 \leq \dot{V}_i \leq \dot{V}_T$. Multiplying both sides of (3.3A) by the positive quantity $4\alpha C^2 f'(v_o)\,\dot{V}_i$ we obtain

$$[2\alpha Cf'(v_o)\dot{V}_i]^2 - 2[2\alpha Cf'(v_o)\dot{V}_i]\left[\frac{1}{\alpha Rf(v_o)} - \frac{1+RG(v_o)}{R}\right] + \frac{4Cf'(v_o)}{Lf(v_o)}(v_o-E_o) > 0 \qquad v_o \geq E_o \tag{3.3B}$$

The left side is a quadratic in $\dot{V}_i$. For $\dot{V}_i=0$ and $\dot{V}_i \to \infty$ it is positive ($v_o \geq E_o$) and for $0<\dot{V}_i<\infty$ it can only be negative at values of v_o such that the quadratic has real and positive roots and when the value of $\dot{V}_i$ lies be-

tween these roots.

Thus let S denote the set of values of $v_o \geq E_o$ for which the roots of (3.3B) are real and positive†; then (3.3A) follows if at each $v_o \in S$ $\dot{V}_i$ is less than the smaller root, i.e., if

$$\dot{V}_i < \frac{1}{2\alpha C f'(v_o)} \left\{ \frac{1}{\alpha R f(v_o)} - \frac{1+RG(v_o)}{R} - \sqrt{\left[\frac{1}{\alpha R f(v_o)} - \frac{1+RG(v_o)}{R} \right]^2 - \frac{4C(v_o-E_o)f'(v_o)}{Lf(v_o)}} \right\} \quad v_o \in S \qquad (3.4)$$

S contains those v_o for which $v_o \geq E_o$ and

$$\frac{1}{\alpha R f(v_o)} - \frac{1+RG(v_o)}{R} \geq \sqrt{\frac{4C(v_o-E_o)f'(v_o)}{Lf(v_o)}}$$

The right side of (3.4) will attain a minimum (positive) value at some $v_o \in S$. On denoting this value of v_o as V_{os} substitution into the right side of (3.4) yields the desired expression for $\dot{V}_T$. To express this result succinctly we define

$$\delta \equiv \sqrt{1 - \frac{\frac{4C}{L}(V_{os}-E_o)\frac{f'(V_{os})}{f(V_{os})}}{\left[\frac{1}{\alpha R f(V_{os})} - \frac{1+RG(V_{os})}{R} \right]^2}} \qquad (3.5)$$

(Note $0<\delta<1$). After simple algebraic manipulation.

$$\dot{V}_T = \frac{V_{os}-E_o}{L} \left[\frac{2}{1+\delta} \right] \left[\frac{R}{1-\alpha f(V_{os})[1+RG(V_{os})]} \right] \qquad (3.6)$$

To recapitulate, we have shown that for inputs, $v_i(t)$, with constant positive slopes satisfying $\frac{dv_i}{dt} = \dot{V}_i < \dot{V}_T$ where $\dot{V}_T$ is given by Equation (3.6), then the output slope will satisfy $\frac{dv_o}{dt} < \alpha \dot{V}_i$ for all values of time. If over any time interval the value of dv_i/dt were to be further reduced to a value below $\dot{V}_i$ the effect if any would be to decrease the maximum value achieved by $\frac{dv_o}{dt}$ so that $\frac{dv_o}{dt} < \alpha \dot{V}_i$ would still be observed. Thus if $v_i(t)$ is composed of a sequence of linear portions for which $\dot{V}_i$ represents the largest slope, then if $\dot{V}_i < \dot{V}_T$ we will still observe $\frac{dv_o}{dt} < \alpha \dot{V}_i$. Carrying this reasoning to the limiting case as $\frac{dv_i}{dt}$ becomes continuous where V_i denotes the max. slope, we still find that $\frac{dv_o}{dt} < \alpha \dot{V}_i$ if $\dot{V}_i < \dot{V}_T$. Thus we observe that for input pulses whose maximum slope fails to exceed the calculated level $\dot{V}_T$, the response will exhibit a maximum slope no greater than α (the prescribed value) times the maximum input slope.

This quantity α represents the degree of attenuation between the input and output signals. For a given input signal whose slope reaches a maximum level of $\dot{V}_i$ we may in a manner similar to the above analysis seek to determine the smallest value of α for which we can guarantee $\frac{dv_o}{dt} < \alpha \dot{V}_i$.

If the resulting value of α is less than unity we know that attenuation is manifest and that the degree of attenuation is at least given by the value of α we have determined.

†For certain choices of network parameters it may occur that S is empty in which case we must conclude that amplification (and thus normal neuristor operation) is impossible. This situation and its relevance to neuristor design will be discussed in a future publication.

IV. RECOVERY TIME

A typical pulse response, $v_o(t)$, terminates with an undershoot followed by a gradual asymptotic recovery of equilibrium as in Fig. 5A. This recovery period in the tail of the pulse corresponds to the refractory period in the neuristor. The associated temporal pulse spacing required for proper neuristor operation can, therefore, be calculated by determining the time required for the voltage during recovery to reach a value from which a new pulse can be readily triggered.

Because recovery is generally a gradual process in the neuristor, one finds in the circuit of Fig. 2 that voltages at adjacent sections differ so slightly during recovery (except for a time interval negligible with respect to the total recovery time at the start of the pulse tail) that we may assume

$$v_n(t) \simeq v_{n-1}(t) \simeq v_{n+1}(t)$$

Applying this to Equaiton (2.1) we find

$$v_i(t) \simeq v_n(t) = v_o(t) \tag{4.1}$$

which on substitution into Equation (2.3) yields

$$\frac{dy_o}{dv_o} \simeq \frac{E_o - v_o}{y_o LC} - \frac{G(v_o)}{C} .$$

Solving for y_o

$$y_o = \frac{E_o - v_o}{LC \dfrac{dy_o}{dv_o} + LG(v_o)} \tag{4.2}$$

The resulting phase plane trajectories corresponding to recovery are typified by the curve of Fig. 5B. Note that y_o changes very slowly except at the very start of recovery. While $\frac{dy_o}{dv_o}$ is initially very large it drops to a small value in a time interval which has been found to be negligibly smaller than the total recovery time. This suggests that an approximate expression for the trajectory might be obtained from Equation (4.2) by neglecting $\frac{dy_o}{dv_o}$.

Specifically, assuming

$$\left|\frac{dy_o}{dv_o}\right| << \left|\frac{G(v_o)}{C}\right|^{\dagger} \tag{4.3}$$

we find the trajectory to be given by

$$y_o(v_o) \simeq \frac{E_o - v_o}{LG(v_o)} \tag{4.4}$$

in which case the time required for $v_o(t)$ to return from some $v=V_a$ to any $v=V_b$ where $V_a \leq V_b < E_o$ (e.g., $V = V_a + 90\% (E_o - V_a)$) is

---

†Numerical analysis of Fig. 2 in a typical example for which $\frac{G(v)}{C} \geq 8.171$ indicates that during recovery the average value of $\frac{dy_o}{dv_o}$ is only .0067.

$$\tau_R = \int_0^{\tau_R} dt = \int_{V_a}^{V_b} \frac{dv_o}{y_o(v_o)} \simeq \int_{V_a}^{V_b} \frac{LG(v_o)}{E_o - v_o}\, dv_o \qquad (4.5)$$

To establish the validity of the above approximations consider an alternate analysis based on a piecewise linearization of the $i(v_o)$ characteristic. Replacing $G(v_o)$ with a sequence of constant conductances G_{Dj} adjusted at time instances $t=t_j$ ($t_o=0$) to maintain G_{Dj} within say 10% of $G(v_o(t))$ and matching values of $v_o=V_{oj}$ ($V_{oo}=V_a$) and $y_o=Y_{oj}$ at $t=t_j$ we obtain the response

$$v_o(t) = E_o + A_j e^{S_{1j}(t-t_j)} + B_j e^{S_{2j}(t-t_j)} \qquad t_j < t \leq t_{j+1} \quad j=0,1,2,\cdots,N \qquad (4.6)$$

where N is selected such that $v_o(t_N) < .9\ E_o + .1\ V_a < v_o(t_{N+1})$ and

$$A_j = \frac{S_{2j}(V_{oj}-E_o) - Y_{oj}}{S_{2j} - S_{1j}}, \qquad B_j = \frac{S_{1j}(E_o-V_{oj}) + Y_{oj}}{S_{2j} - S_{1j}},$$

$$S_{1j} = -\left(\frac{G_{Dj}}{2C}\right) + \sqrt{\left(\frac{G_{Dj}}{2C}\right)^2 - \frac{1}{LC}}, \qquad S_{2j} = -\left(\frac{G_{Dj}}{2C}\right) - \sqrt{\left(\frac{G_{Dj}}{2C}\right)^2 - \frac{1}{LC}}$$

The value of τ_R is then the value of $t>t_N$ for which $v_o(\tau_R) = .9\ E_o + .1\ V_a$.

An explicit formula for τ_R will, however, require that we introduce a further approximation. Noting that $G(v_o)$ generally attains relatively large values for $v_o<E_o$, we may assume

$$\left| \frac{G(v_o)}{C} \right| >> \left| \frac{1}{LG(v_o)} \right| \qquad (4.7A)$$

In this case S_{1j} and S_{2j} are real and subject to a dominant pole approximation

$$\left|S_{ij}\right| << \left|S_{2j}\right| \qquad (4.7B)$$

which allows (4.6) to be rewritten as

$$v_o(t) = E_o + (V_{oj}-E_o)\, e^{-\frac{t-t_j}{LG_{Dj}}} \qquad t_j < t \leq t_{j+1} \qquad (4.8)$$

The resulting expression for τ_R is

$$\tau_R = (\tau_R - t_N) + \sum_{j=0}^{N-1} (t_{j+1}-t_j)$$

$$\tau_R = LG_{DN} \ell n\ 10 \left[\frac{E_o-V_{oN}}{E_o-V_a}\right] + \sum_{j=0}^{N-1} L\, G_{Dj} \ell n \left[\frac{E_o - V_{oj}}{E_o - V_{o,j+1}}\right] \qquad (4.9)$$

This result is not unrelated to Equation (4.5). If the same linearized form of $G(v_o)$ is substituted into (4.5) the result will be (4.9). Conversely if we improve the linearization (i.e., require G_{Dj} and $G(v_o)$ to differ by less than 10%) to the limit as $N\to\infty$ and $G_{Dj}\to G_{D,j+1}$ then Equation (4.9) reduces to (4.5). This actually suggests that approximation (4.4) is in effect equivalent to the dominant pole approximation, (4.7B), which suggestion is

supported by the similarity between (4.3) and (4.7A). Therefore, while retaining the simplicity lacking in (4.6), Equation (4.5) represents a limiting form of (4.9) (obtained by removal of the linearization) and should be uniformly more accurate than (4.9).

V. EXAMPLES

A thorough discussion of the application of the above results to neuristor analysis is not intended in this paper, but we can demonstrate how this is to be accomplished by example. For this purpose typical element values for Figs. 1 and 2 have been selected and a typical tunnel diode approximated[7]. After voltage and current scaling to obtain for the diode $I_p=1$ at $V_p=.1$, and $I_v=.05$ at $V_v=.622$ and time scaling in a ratio of 1 microsecond to 1 second these parameter values are

$$R=.513,\ R_1=1.026,\ C=1.167,\ L=4.350,\ E_o=.050$$

$$i(v_o) = 10v_o \exp(1-10v_o) + \exp(9.2v_o-9.2)$$

$$G(v_o) = 10(1-10v_o) \exp(1-10v_o) + 9.2 \exp(9.2v_o-9.2)$$

Plugging these values into the right side of (3.4) with $\alpha=1$ and $k=.4$[†] we find $V_{os}=.124$ and $\dot{V}_T=.00886$. This means that when a signal $v_i(t)$ for which $\frac{dv_i}{dt} < .00886$ is applied to the circuit of Fig. 1 the response slope, $\frac{dv_o}{dt}$, will at all times satisfy $\frac{dv_o}{dt} < \frac{dv_i}{dt}$.

Using a packaged predictor-corrector scheme (HPCG from IBM'S scientific subroutine package) to provide a computer simulation of the corresponding neuristor (Fig. 2) it is observed that inputs $v_{IN}(t)$ of constant slope $\dot{V}_{IN}$ (sustained as long as necessary for $\frac{dv_1}{dt}$ to reach a maximum) are subject to a threshold level of about .0177 v/sec.

Even for slightly higher values of $\dot{V}_{IN}$ it is found that during the time that $\frac{dv_1}{dt}$ is rising towards its maximum level, it is safe to assume (at least on the average) that $\frac{dv_2}{dt} < .1\ \dot{V}_{IN}$. Applying this to Equation (2.1A) with n=1 we can use the analysis of Fig. 1 (Section III) to predict a lower estimate of the threshold in Fig. 2 as follows. Equation (2.1A) yields

$$\frac{dv_i}{dt} = \frac{1}{2}\left[\frac{dv_{IN}}{dt} + \frac{dv_2}{dt}\right] < \frac{1.1}{2}\ \frac{dv_{IN}}{dt} = \frac{1.1}{2}\ \dot{V}_{IN} \qquad (5.1)$$

and the analysis of Fig. 1 predicts $\frac{dv_1}{dt}\ \left(\equiv \frac{dv_o}{dt}\right) < \text{MAX}\ \frac{dv_2}{dt}$ pointwise if $\frac{dv_i}{dt} < .00886$ pointwise. Since $\dot{V}_{IN} < \frac{2}{1.1}\ (.00886)$ in (5.1) implies $\frac{dv_i}{dt} < .00886$, we conclude that threshold in Fig. 2 will not be broken if

---

†This value was selected after repeating calculations of $\dot{V}_T$ using k=.25, .4, .5, .6, .7, .8 and .9. For each value of k a list of values of the right side of (3.4) was computed with v_o ranging upwards from E_o in steps of 2^{-9}. At each value of k the least value listed is the corresponding value of $\dot{V}_T$. The choice k=.4 led to the highest (best) prediction of $\dot{V}_T$.

$\dot{V}_{1N} \leq \frac{2}{1.1}(.00886)=.01611$ v/sec. This result is within 10% of the actual threshold level and as claimed is a lower estimate of the actual value. Further discussion of this lower bound along with an analysis leading to an upper bound will be the subject of a future paper.

For the same neuristor the recovery time has been studied both by simulation and via Equation (4.5). It is observed that the propagating pulse in the simulation exhibits an undershoot of $V_a=-.0069$ volts and returns to $V_b=V_a+.9(E_o-V_a)=.0443$ volts in 154.99 seconds. The predicted value obtained from Equation (4.5) using $G(v_o)\approx 10(1-10v_o)\exp(1-10v_o)$ (for $v<E_o$ the additional term contributes less than 1% of the total value of $G(v_o)$) is

$$\tau_R \approx \int_{V_a}^{V_b} \frac{LG(v_o)}{E_o-v_o}\,dv_o \approx \int_{-.0069}^{.0443} \frac{435(.1-v_o)\exp(1-10v_o)}{.05-v_o}\,dv_o = 155.1 \text{ sec.}$$

which deviates from the simulation by .07%.

VI. SUMMARY

This paper has reported an analysis of pulse attenuation and recovery time in the nonlinear active circuit of Fig. 1 which is an elementary section of the neuristor of Fig. 2. Numerical examples demonstrated the applicability of this analysis to the estimation of threshold and recovery time in the neuristor.

The analysis of Fig. 1 was based on geometrical considerations in the phase plane which allowed us to avoid actually solving the nonlinear network equations. For example in studying pulse attenuation we constructed a "test curve" of convenient form, and by simply employing algebraic manipulations we were then able to show that the trajectory must lie below the test curve and thus that attenuation by at least a given factor, α, must occur whenever the applied voltage fails to exhibit slopes in excess of a value $\dot{V}_T$ given by Equation (3.6). For the estimation of recovery time it was found possible to employ an approximate expression for the actual trajectory to arrive directly at Equation (4.5). This formula was shown to offer a better compromise between simplicity and accuracy when compared against the results of a solution obtained via piecewise linear approximation.

By extending the techniques reported here it should be possible to evaluate the performance of a variety of neuristor models. For instance Noguchi et al [3] proposed a model whose elementary section is identical to Fig. 1 if we replace the resistance, R, by an inductance. This model, however responds to pulse amplitude whereas the model proposed by Nagumo et al[2] (Fig. 2) responds to pulse slope. It would be of value to know if this amplitude sensitivity, which more closely approximates nerve axon behavior, offers any advantage over slope sensitivity. Thus the results reported here may ultimately be used in comparing performance as well as in analyzing and designing neuristor circuits.

REFERENCES

1. H.D. Crane, "Neuristor-A Novel Device and System Concept," Proc. IRE, vol. 50, No. 10, 1962.

2. Nagumo, Arimoto and Yoshizawa, "An Active Pulse Transmission Line Simulating Nerve Axon," Proc. IRE, vol. 50, No. 10, 1962.

3. Noguchi, Kumagaya, Tsutiya, Nishizawa, and Oizumi, "Fundamental Characteristics of the Electronic Neuristor Circuit," Electronics and Communications in Japan, vol. 46, No. 11, 1963.

4. Hayasaka, and Nishizawa, "Pulse Characteristics of the Distributed Esaki Diode," Electronics and Communications in Japan, vol. 49, No. 8, 1966.

5. Berestovskiy, G.N., "Study of Single Electric Model of Neuristor," Radio Engineering and Electronic Physics, vol. 8, No. 11, 1963.

6. N. Minorsky, Introduction to Nonlinear Mechanics, Chapters I-III, J.W. Edwards, Ann Arbor, 1947.

7. Ferendeci, and Ko, "A Two Term Analytical Approximation of Tunnel Diode Static Characteristics," Proc. IRE, vol. 50, No. 8, 1962.

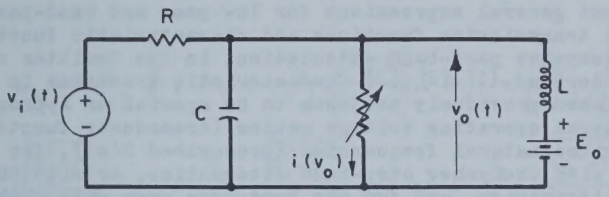


FIG. 1

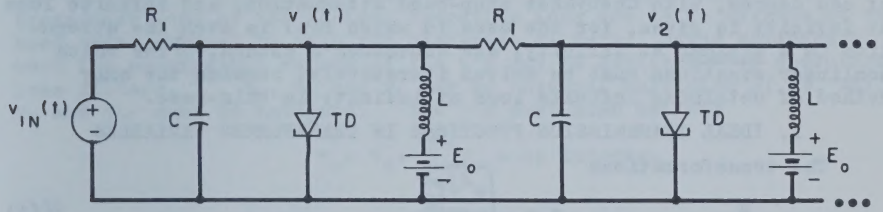


FIG. 2

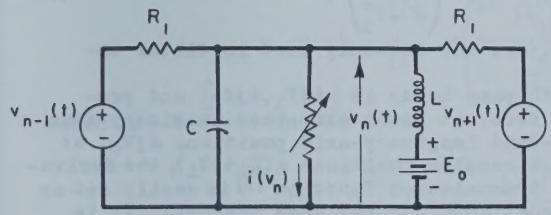


FIG. 3

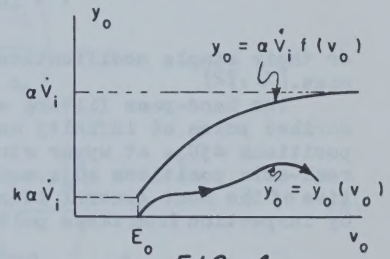


FIG. 4

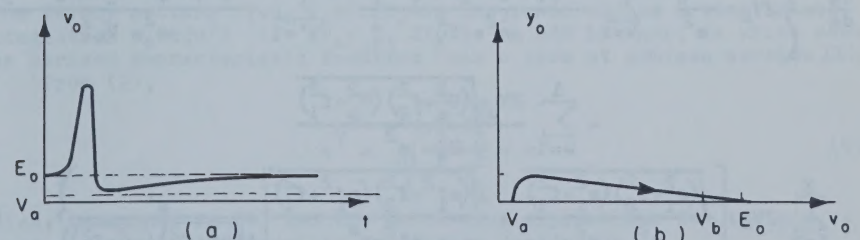


FIG. 5

GENERAL EXPRESSIONS FOR IDEAL TRANSMISSION FUNCTIONS, CHARACTERISTIC FUNCTIONS AND OPERATING VOLTAGE RATIOS FOR FILTER DESIGN USING TRANSFORMED VARIABLES

M. SABLATASH
Dept. of Electrical Engineering
The University of Toronto
Toronto, Canada

ABSTRACT

The most general expressions for low-pass and band-pass generalized ideal transmission functions and characteristic functions yielding Chebyshev pass-band attenuation, in the familiar z- and Y-planes are derived.[1],[2],[3] Characteristic functions in the z-plane published previously are seen to be special or approximating cases. General operating voltage ratios (transducer functions), H(s), with prescribed natural frequencies (prescribed E(s)), for the low-pass cases with Chebyshev stop-band attenuation, or with Chebyshev pass-band attenuation, and for the band-pass case with Chebyshev pass-band attenuation are found in closed-form. Although a closed-form general expression for H(s)of a low-pass filter with prescribed E(s) of odd degree, with Chebyshev stop-band attenuation, and infinite loss at infinity is given, for the case in which E(s) is even the attenuation is a minimum at infinity, and Unbehauen's method,[2] for which nonlinear equations must be solved iteratively, remains the only method of obtaining infinite loss at infinity in this case.

I. IDEAL TRANSMISSION FUNCTIONS IN TRANSFORMED VARIABLES

The transformations

$$z = \sqrt{\frac{s^2+f_2^2}{s^2+f_1^2}} \tag{1}$$

and

$$Y = \ln z = \tfrac{1}{2} \ln \left(\frac{s^2+f_2^2}{s^2+f_1^2} \right), \tag{2}$$

or their simple modifications, are now commonly used in filter design.[1],[2]

For band-pass filters with pass bands in $[\pm jf_1, \pm jf_2]$ and prescribed poles at infinity and zero, at lower stop-band imaginary-axis positions $\pm jU_{\nu'}$, at upper stop-band imaginary-axis positions $\pm jU_{\nu}$, at real-axis positions $\pm U_{\mu}$, and at complex positions $\pm(U_i \pm jV_i)$, the derivative of the most general ideal transmission function[3] is easily set up by inspection from these pole positions and residues thereat. It is

$$\frac{dw}{ds} = \left\{ 2Y_\infty + \frac{2Y_o}{s^2} + \sum_{\nu=1}^{h_1} \frac{2Y_\nu \sqrt{(V_\nu^2-f_2^2)(V_\nu^2-f_1^2)}}{s^2+V_\nu^2} + \sum_{\nu'=1}^{h_2} \frac{2Y_{\nu'} \sqrt{(U_{\nu'}^2-f_2^2)(U_{\nu'}^2-f_1^2)}}{s^2+U_{\nu'}^2} \right.$$

$$+ \sum_{\mu=1}^{l} \frac{2Y_\mu \sqrt{(U_\mu^2+f_2^2)(U_\mu^2+f_1^2)}}{U_\mu^2 - s^2}$$

$$\left. + \sum_{i=1}^{g} 2Y_i \left[\frac{\sqrt{(s_i^2+f_2^2)(s_i^2+f_1^2)}}{s_i^2 - s^2} + \frac{\sqrt{(s_i^{*2}+f_2^2)(s_i^{*2}+f_1^2)}}{s_i^{*2}-s^2} \right] \frac{s}{\sqrt{(s^2+f_2^2)(s^2+f_1^2)}} \right\}. \tag{3}$$

With $\lambda^2=(s^2+f_1^2)/(f_2^2-f_1^2)$ used to facilitate integration, and substituting (1) in the result, one finds the most general band-pass ideal transmission function,

$$w(z) = |\gamma_\infty| \ln\left(\frac{z+m_\infty}{z-m_\infty}\right) + |\gamma_o| \ln\left(\frac{z+m_o}{z-m_o}\right) + \sum_{\nu=1}^{h_1} |\gamma_\nu| \ln\left(\frac{z+m_\nu}{z-m_\nu}\right)$$

$$+ \sum_{\nu'=1}^{h_2} |\gamma_{\nu'}| \ln\left(\frac{z+m_{\nu'}}{z-m_{\nu'}}\right) + \sum_{\mu=1}^{l} |\gamma_\mu| \ln\left(\frac{z+m_\mu}{z-m_\mu}\right) + \sum_{i=1}^{g} |\gamma_i| \ln\left[\frac{(z+m_i)(z+m_i^*)}{(z-m_i)(z-m_i^*)}\right], \quad (4)$$

where

$$m_\infty = \frac{\gamma_\infty}{|\gamma_\infty|}, \quad m_o = \frac{\gamma_o}{|\gamma_o|}\frac{1}{f_1}, \quad m_\nu = \frac{\gamma_\nu}{|\gamma_\nu|}\sqrt{\frac{v_\nu^2-f_2^2}{v_\nu^2-f_1^2}}, \quad m_{\nu'} = \frac{\gamma_{\nu'}}{|\gamma_{\nu'}|}\sqrt{\frac{f_2^2-U_\nu^2}{f_1^2-U_\nu^2}},$$

$$m_\mu = \frac{\gamma_\mu}{|\gamma_\mu|}\sqrt{\frac{U_\mu^2+f_2^2}{U_\mu^2+f_1^2}}, \quad m_i = \frac{\gamma_i}{|\gamma_i|}\sqrt{\frac{s_i^2+f_2^2}{s_i^2+f_1^2}} \quad \text{and} \quad m_i^* = \frac{\gamma_i}{|\gamma_i|}\sqrt{\frac{s_i^{*2}+f_2^2}{s_i^{*2}+f_1^2}}. \quad (5)$$

The constant of integration must be zero if the magnitudes of the transducer loss of the derived operating loss are to be maximum at the band ends, as must be true for band-pass filters. Using Watanabe's work the permissible values of the γ's may be deduced.[3] For operating loss filters, γ_∞, γ_o, γ_μ and γ_i may be integral multiples of $\pm\frac{1}{2}$, while γ_ν and $\gamma_{\nu'}$ must be integers. The γ's must also satisfy

$$\gamma_\infty + \gamma_o + \sum_{\mu=1}^{l} \gamma_\mu = \text{an integer.}$$

With

$$N = 2|\gamma_\infty| + 2|\gamma_o| + \sum_{\nu=1}^{h_1} 2|\gamma_\nu| + \sum_{\nu'=1}^{h_2} 2|\gamma_{\nu'}| + \sum_{\mu=1}^{l} 2|\gamma_\mu| + \sum_{i=1}^{g} 4|\gamma_i|, \quad (6)$$

and the pole positions in the z-plane denoted by $z=z_i$,

$$w(z) = \sum_{i=1}^{N/2} \ln\left(\frac{z+\frac{\gamma_i}{|\gamma_i|}z_i}{z-\frac{\gamma_i}{|\gamma_i|}z_i}\right), \quad (7)$$

which may be cast in the form

$$w(z) = \sum_{i=1}^{N/2} \ln\left(\frac{z^2+\frac{2\gamma_i}{|\gamma_i|}z_i z+z_i^2}{z^2-z_i^2}\right). \quad (8)$$

The most general low-pass ideal transmission function is obtained from (4) by setting $f_1=0$, $\gamma_o=0$, $h_2=0$, $h=h_1$, and adding a constant of integration $w_o=\pm j\pi/2$ if $2\gamma_\infty+\sum_1 2\gamma_\mu$ is an odd integer, in which case the derived characteristic function has a zero at $s=0$ (see section II).

From (2),

$$e^\gamma = \sqrt{\frac{s^2+f_2^2}{s^2+f_1^2}} = z \quad (9)$$

which, when used in (4) and the succeeding expressions for the m's, yields (with added subscripts r denoting residues)

$$w(\gamma) = |\gamma_{\infty r}| \ln\left[\frac{e^{\gamma}+m_{\infty}}{e^{\gamma}-m_{\infty}}\right] + |\gamma_{or}| \ln\left[\frac{e^{\gamma}+m_{o}}{e^{\gamma}-m_{o}}\right] + \sum_{\nu=1}^{h_1} |\gamma_{\nu r}| \ln\left[\frac{e^{\gamma}+m_{\nu}}{e^{\gamma}-m_{\nu}}\right]$$

$$+ \sum_{\nu'=1}^{h_2} |\gamma_{\nu' r}| \ln\left[\frac{e^{\gamma}+m_{\nu'}}{e^{\gamma}-m_{\nu'}}\right] + \sum_{\mu=1}^{l} |\gamma_{\mu r}| \ln\left[\frac{e^{\gamma}+m_{\mu}}{e^{\gamma}-m_{\mu}}\right] + \sum_{i=1}^{g} |\gamma_{ir}| \ln\left[\frac{(e^{\gamma}+m_i)(e^{\gamma}+m_i^*)}{(e^{\gamma}-m_i)(e^{\gamma}-m_i^*)}\right], \quad (10)$$

where

$$m_{\infty} = \frac{\gamma_{\infty r}}{|\gamma_{\infty r}|} e^{\gamma_{\infty}} = \frac{\gamma_{\infty r}}{|\gamma_{\infty r}|} e^{0} = \frac{\gamma_{\infty r}}{|\gamma_{\infty r}|}, \qquad m_{o} = \frac{\gamma_{or}}{|\gamma_{or}|} e^{\gamma_o} = \frac{\gamma_{or}}{|\gamma_{or}|} \frac{f_2}{f_1},$$

$$m_{\nu'} = \frac{\gamma_{\nu' r}}{|\gamma_{\nu' r}|} e^{\gamma_{\nu'}} = \frac{\gamma_{\nu' r}}{|\gamma_{\nu' r}|} \sqrt{\frac{f_2^2-U_{\nu'}^2}{f_1^2-U_{\nu'}^2}}, \qquad m_{\nu} = \frac{\gamma_{\nu r}}{|\gamma_{\nu r}|} e^{\gamma_{\nu}} = \frac{\gamma_{\nu r}}{|\gamma_{\nu r}|} \sqrt{\frac{V_{\nu}^2-f_2^2}{V_{\nu}^2-f_1^2}},$$

$$m_{\mu} = \frac{\gamma_{\mu r}}{|\gamma_{\mu r}|} e^{\gamma_{\mu}} = \frac{\gamma_{\mu r}}{|\gamma_{\mu r}|} \sqrt{\frac{U_{\mu}^2+f_2^2}{U_{\mu}^2+f_1^2}}, \qquad m_i = \frac{\gamma_{ir}}{|\gamma_{ir}|} e^{\gamma_i} = \frac{\gamma_{ir}}{|\gamma_{ir}|} \sqrt{\frac{s_i^2+f_2^2}{s_i^2+f_1^2}},$$

and

$$m_i^* = \frac{\gamma_{ir}}{|\gamma_{ir}|} e^{\gamma_i^*} = \frac{\gamma_{ir}}{|\gamma_{ir}|} \sqrt{\frac{s_i^{*2}+f_2^2}{s_i^{*2}+f_1^2}}.$$

Equation (10) is easily cast into

$$w(\gamma) = \gamma_{\infty r} \ln\left[\frac{e^{\gamma}+e^{\gamma_{\infty}}}{e^{\gamma}-e^{\gamma_{\infty}}}\right] + \gamma_{or} \ln\left[\frac{e^{\gamma}+e^{\gamma_o}}{e^{\gamma}-e^{\gamma_o}}\right] + \sum_{\nu=1}^{h_1} \gamma_{\nu r} \ln\left[\frac{e^{\gamma}+e^{\gamma_{\nu}}}{e^{\gamma}-e^{\gamma_{\nu}}}\right]$$

$$+ \sum_{\nu'=1}^{h_2} \gamma_{\nu' r} \ln\left[\frac{e^{\gamma}+e^{\gamma_{\nu'}}}{e^{\gamma}-e^{\gamma_{\nu'}}}\right] + \sum_{\mu=1}^{l} \gamma_{\mu r} \ln\left[\frac{e^{\gamma}+e^{\gamma_{\mu}}}{e^{\gamma}-e^{\gamma_{\mu}}}\right] + \sum_{i=1}^{g} \gamma_{ir} \ln\left[\frac{(e^{\gamma}+e^{\gamma_i})(e^{\gamma}+e^{\gamma_i^*})}{(e^{\gamma}-e^{\gamma_i})(e^{\gamma}-e^{\gamma_i^*})}\right]. \quad (11)$$

Simple algebraic steps yield

$$w(\gamma) = \gamma_{\infty r} \ln \coth\left(\frac{\gamma-\gamma_{\infty}}{2}\right) + \gamma_{or} \ln \coth\left(\frac{\gamma-\gamma_o}{2}\right) + \sum_{\nu=1}^{h_1} \gamma_{\nu r} \ln \coth\left(\frac{\gamma-\gamma_{\nu}}{2}\right)$$

$$+ \sum_{\nu'=1}^{h_2} \gamma_{\nu' r} \ln \coth\left(\frac{\gamma-\gamma_{\nu'}}{2}\right) + \sum_{\mu=1}^{l} \gamma_{\mu r} \ln \coth\left(\frac{\gamma-\gamma_{\mu}}{2}\right)$$

$$+ \sum_{i=1}^{g} \gamma_{ir} \ln\left[\coth\left(\frac{\gamma-\gamma_i}{2}\right) \coth\left(\frac{\gamma-\gamma_i^*}{2}\right)\right]. \quad (12)$$

Even more compactly,

$$w(\gamma) = \sum_{i=1}^{N/2} \gamma_{ir} \ln \coth\left(\frac{\gamma-\gamma_i}{2}\right), \quad (13)$$

wherein N is given by (6), and the γ_i are the pole positions in the γ-plane. The attenuation in the γ-plane is obtained by taking the magnitude of the arguments of the coth in (12) and (13). The low-pass

case is obtained by setting $f_1=0$, $\gamma_{or}=0$, and $h_2=0$. In this way Orchard and Temes' equations (26), (27), (64) and (80) are easily obtained.[1] Note that the residues in (12) and (13) may be positive or negative, in general, and that complex and real poles are included.

II. GENERAL EXPRESSIONS FOR CHARACTERISTIC FUNCTIONS FOR FILTER DESIGN USING TRANSFORMED VARIABLES

The generalized characteristic function (with pass-band maxima and minima $\pm\epsilon$) in the z-plane is

$$\varphi(z) = \epsilon \cosh\left[w(z)\right] . \tag{14}$$

In the low-pass case (4) is easily cast in the form

$$w(z) = |\gamma_\infty| \ln\left(\frac{z^2+2m_\infty z+m_\infty^2}{z^2-m_\infty^2}\right) + \sum_{\nu=1}^{h} |\gamma_\nu| \ln\left(\frac{z^2+2m_\nu z+m_\nu^2}{z^2-m_\nu^2}\right) + \sum_{\mu=1}^{l} |\gamma_\mu| \ln\left(\frac{z^2+2m_\mu z+m_\mu^2}{z^2-m_\mu^2}\right)$$

$$+ \sum_{i=1}^{g} |\gamma_i| \ln\left[\frac{\left(z^2+2m_i z+m_i^2\right)\left(z^2+2m_i^* z+m_i^{*2}\right)}{\left(z^2-m_i^2\right)\left(z^2-m_i^{*2}\right)}\right] + w_o . \tag{15}$$

Similarly,

$$-w(z) = |\gamma_\infty| \ln\left(\frac{z^2-2m_\infty z+m_\infty^2}{z^2-m_\infty^2}\right) + \sum_{\nu=1}^{h} |\gamma_\nu| \ln\left(\frac{z^2-2m_\nu z+m_\nu^2}{z^2-m_\nu^2}\right) + \sum_{\mu=1}^{l} |\gamma_\mu| \ln\left(\frac{z^2-2m_\mu z+m_\mu^2}{z^2-m_\mu^2}\right)$$

$$+ \sum_{i=1}^{g} |\gamma_i| \ln\left[\frac{\left(z^2-2m_i z+m_i^2\right)\left(z^2-2m_i^* z+m_i^{*2}\right)}{\left(z^2-m_i^2\right)\left(z^2-m_i^{*2}\right)}\right] - w_o . \tag{16}$$

Hence, from (14), (15) and (16),

$$\varphi(z) = \frac{\epsilon}{2}\left[e^{w(z)}+e^{-w(z)}\right] = \frac{\epsilon}{2}\left\{\left[\frac{\left(z^2+2m_\infty z+m_\infty^2\right)^{|\gamma_\infty|}}{\left(z^2-1\right)^{|\gamma_\infty|}} \prod_{\nu=1}^{h} \frac{\left(z^2+2m_\nu z+m_\nu^2\right)^{|\gamma_\nu|}}{\left(z^2-m_\nu^2\right)^{|\gamma_\nu|}}\right.\right.$$

$$\left. \cdot \prod_{\mu=1}^{l} \frac{\left(z^2+2m_\mu z+m_\mu^2\right)^{|\gamma_\mu|}}{\left(z^2-m_\mu^2\right)^{|\gamma_\mu|}} \prod_{i=1}^{g} \frac{\left(z^2+2m_i z+m_i^2\right)^{|\gamma_i|}\left(z^2+2m_i^* z+m_i^{*2}\right)^{|\gamma_i|}}{\left(z^2-m_i^2\right)^{|\gamma_i|}\left(z^2-m_i^{*2}\right)^{|\gamma_i|}}\right] e^{w_o}$$

$$+ \left[\frac{\left(z^2-2m_\infty z+m_\infty^2\right)^{|\gamma_\infty|}}{\left(z^2-1\right)^{|\gamma_\infty|}} \prod_{\nu=1}^{h} \frac{\left(z^2-2m_\nu z+m_\nu^2\right)^{|\gamma_\nu|}}{\left(z^2-m_\nu^2\right)^{|\gamma_\nu|}} \prod_{\mu=1}^{l} \frac{\left(z^2-2m_\mu z+m_\mu^2\right)^{|\gamma_\mu|}}{\left(z^2-m_\mu^2\right)^{|\gamma_\mu|}}\right.$$

$$\left.\left. \cdot \prod_{i=1}^{g} \frac{\left(z^2-2m_i z+m_i^2\right)^{|\gamma_i|}\left(z^2-2m_i^* z+m_i^{*2}\right)^{|\gamma_i|}}{\left(z^2-m_i^2\right)^{|\gamma_i|}\left(z^2-m_i^{*2}\right)^{|\gamma_i|}}\right] e^{-w_o}\right\} . \tag{17}$$

If $\varphi(s=0)=0$, $2\gamma_\infty+\sum_{\mu=1}^{l} 2\gamma_\mu$ is an odd integer and $w_o=\pm j\frac{\pi}{2}$, and it is useful to define

$$\varphi_1(s) = j\varphi(s). \tag{18}$$

If $\varphi(s=0)=\pm\epsilon$, $2\gamma_\infty+\sum_{\mu=1}^{l} 2\gamma_\mu$ is an even integer and $w_o=0$,

$$\varphi_1(s) = \varphi(s). \tag{19}$$

Then

$$\varphi_1(z) = \frac{\varepsilon \mathrm{Ev}\left[\left(z^2+2m_\infty z+m_\infty^2\right)^{|\gamma_\infty|}\prod_{\nu=1}^{h}\left(z^2+2m_\nu z+m_\nu^2\right)^{|\gamma_\nu|}\prod_{\mu=1}^{l}\left(z^2+2m_\mu z+m_\mu^2\right)^{|\gamma_\mu|}\cdot\prod_{i=1}^{g}\left(z^2+2m_i z+m_i^2\right)^{|\gamma_i|}\left(z^2+2m_i^* z+m_i^{*2}\right)^{|\gamma_i|}\right]}{\left(z^2-1\right)^{|\gamma_\infty|}\prod_{\nu=1}^{h}\left(z^2-m_\nu^2\right)^{|\gamma_\nu|}\prod_{\mu=1}^{l}\left(z^2-m_\mu^2\right)^{|\gamma_\mu|}\prod_{i=1}^{g}\left[\left(z^2-m_i^2\right)\left(z^2-m_i^{*2}\right)\right]^{|\gamma_i|}} \quad (20)$$

The most general perfectly equal-ripple band-pass generalized characteristic function is

$$\varphi(z) = \frac{\varepsilon \mathrm{Ev}\left[\left(z+m_\infty\right)^{2|\gamma_\infty|}\left(z+m_o\right)^{2|\gamma_o|}\prod_{\nu=1}^{h_1}\left(z+m_\nu\right)^{2|\gamma_\nu|}\prod_{\nu'=1}^{h_2}\left(z+m_{\nu'}\right)^{2|\gamma_{\nu'}|}\prod_{\mu=1}^{l}\left(z+m_\mu\right)^{2|\gamma_\mu|}\prod_{i=1}^{g}\left(z+m_i\right)^{2|\gamma_i|}\left(z+m_i^*\right)^{2|\gamma_i|}\right]}{\left(z^2-1\right)^{|\gamma_\infty|}\left(z^2-m_o^2\right)^{|\gamma_o|}\prod_{\nu=1}^{h_1}\left(z^2-m_\nu^2\right)^{|\gamma_\nu|}\prod_{\nu'=1}^{h_2}\left(z^2-m_{\nu'}^2\right)^{|\gamma_{\nu'}|}\prod_{\mu=1}^{l}\left(z^2-m_\mu^2\right)^{|\gamma_\mu|}\cdot\prod_{i=1}^{g}\left(z^2-m_i^2\right)^{|\gamma_i|}\left(z^2-m_i^{*2}\right)^{|\gamma_i|}} \quad (21)$$

From (21), at z=0, the mapping of the pass-band edge $s=\pm jf_2$,

$$\varphi(z=0) = \frac{1}{\sqrt{10^{0.1A_{RL}}-1}} = \varepsilon(-1)^{|\gamma_\infty|+|\gamma_o|+\sum_{\nu=1}^{h_1}|\gamma_\nu|+\sum_{\nu'=1}^{h_2}|\gamma_{\nu'}|+\sum_{\mu=1}^{l}|\gamma_\mu|+\sum_{i=1}^{g}2|\gamma_i|} \quad (22)$$

Expressions for $\varphi(\gamma)$ (not used in design yet) may be obtained by setting $z=e^{\gamma}$. It should be noted that all previously known low-pass and band-pass characteristic functions with perfectly-ripple pass-bands are included in (20) and (21). Approximations are easily made by methods shown in the writer's earlier work and in Orchard and Temes' paper.[4], [1]

III. GENERAL EXPRESSIONS FOR OPERATING VOLTAGE RATIOS WITH PRESCRIBED NATURAL FREQUENCIES (NUMERATOR)

The problem of obtaining Chebyshev pass-band or stop-band attenuation when the natural frequencies have been prescribed has been considered by a number of previous workers.[2],[5],[6],[7] General solutions (closed-form for all low-pass cases and for the band-pass case with Chebyshev pass-band attenuation), in convenient form for computation, are given here. Denoting the operating voltage ratio, or transducer function, by $H(s)=E(s)/P(s)$, the problem of finding P(s), given E(s), so as to obtain equal minima of stop-band operating attenuation in $\omega_S \leq \omega \leq \infty$, is solved as follows, for the degree, N, of E(s) even or odd. One notes that

$$[S(s)]^{-1} = \frac{P(s)}{\sqrt{E(s)E(-s)}} \quad (23)$$

must have equal-ripple deviation from zero in $\omega_S \leq \omega \leq \infty$. With real zeros of E(s) at $s=u_\mu$ and complex zeros at $s=s_i$ and s_i^*,

$$\frac{dw}{ds} = \left[\sum_{\mu=1}^{l} \frac{2\gamma_\mu s\sqrt{u_\mu^2+\omega_s^2}}{u_\mu^2-s^2} + \sum_{i=1}^{g} 2\gamma_i\left(\frac{s\sqrt{s_i^2+\omega_s^2}}{s_i^2-s^2} + \frac{s\sqrt{s_i^{*2}+\omega_s^2}}{s_i^{*2}-s^2}\right)\right]\frac{1}{\sqrt{s^2+\omega_s^2}}, \tag{24}$$

from which

$$w(s)=\sum_{\mu=1}^{l}\gamma_\mu\cosh^{-1}\left(\frac{u_\mu^2+2\omega_s^2+s^2}{u_\mu^2-s^2}\right)+\sum_{i=1}^{g}2\gamma_i\cosh^{-1}\left(\frac{\sqrt{(\omega_s^2+s_i^2)(\omega_s^2+s_i^{*2})}+\omega_s^2+s^2}{\sqrt{(s_i^2-s^2)(s_i^{*2}-s^2)}}\right)+w_o. \tag{25}$$

Using $\cosh^{-1}(s)=\ln s+\sqrt{s^2+1}$, (25) can be cast in the form

$$w(s) = \sum_{\mu=1}^{l}|\gamma_\mu|\ln\left[\frac{u_\mu^2+2\omega_s^2+s^2+2\frac{\gamma_\mu}{|\gamma_\mu|}\sqrt{(s^2+\omega_s^2)(u_\mu^2+\omega_s^2)}}{u_\mu^2-s^2}\right]$$

$$+\sum_{i=1}^{g}2|\gamma_i|\ln\left[\frac{\sqrt{(\omega_s^2+s_i^2)(\omega_s^2+s_i^{*2})}+\omega_s^2+s^2+\frac{\gamma_i}{|\gamma_i|}\left[\sqrt{\omega_s^2+s_i^2}+\sqrt{\omega_s^2+s_i^{*2}}\right]\sqrt{s^2+\omega_s^2}}{\sqrt{(s_i^2-s^2)(s_i^{*2}-s^2)}}\right]+w_o. \tag{26}$$

In order that $\cosh[w(s)]$ be zero for $s=\infty$, $w(\infty)$ must be an odd integral multiple of $j\pi/2$. From (26)

$$w(\infty) = j\pi\left[\sum_{\mu=1}^{l}|\gamma_\mu|+\sum_{i=1}^{g}2|\gamma_i|\right]+w_o. \tag{27}$$

Therefore if $\left[\sum_{\mu=1}^{l}|\gamma_\mu|+\sum_{i=1}^{g}2|\gamma_i|\right]$ is an integer, w_o must be $\pm j\frac{\pi}{2}$, but if $\left[\sum_{\mu=1}^{l}|\gamma_\mu|+\sum_{i=1}^{g}2|\gamma_i|\right]$ is an odd multiple of ½, $w_o=0$. It will be shown that for both cases w_o must be zero. To obtain the form of (23), consider all zeros of E(s) to be distinct even though they may not be, and set all γ_μ and γ_i equal to ±½ (usually +½), and calculate $e^{-\alpha_m}\cosh[w(s)]$, where α_m, the minimum of the stop-band attenuation, is easily calculated by setting $s=0$, where the attenuation must be zero and $H(0)=1$. This yields

$$H(s)^{-1}= e^{-\alpha_m}\frac{\prod_{\mu=1}^{l}\left[u_\mu^2+2\omega_s^2+s^2\pm2\sqrt{(s^2+\omega_s^2)(u_\mu^2+\omega_s^2)}\right]^{1/2}\times\prod_{i=1}^{g}\left[\sqrt{(\omega_s^2+s_i^2)(\omega_s^2+s_i^{*2})}+\omega_s^2+s^2\pm\left(\sqrt{\omega_s^2+s_i^2}+\sqrt{\omega_s^2+s_i^{*2}}\right)\sqrt{s^2+\omega_s^2}\right]e^{w_o}+\prod_{\mu=1}^{l}\left[u_\mu^2+2\omega_s^2+s^2\mp2\sqrt{(s^2+\omega_s^2)(u_\mu^2+\omega_s^2)}\right]^{1/2}\times\prod_{i=1}^{g}\left[\sqrt{(\omega_s^2+s_i^2)(\omega_s^2+s_i^{*2})}+\omega_s^2+s^2\mp\left(\sqrt{\omega_s^2+s_i^2}+\sqrt{\omega_s^2+s_i^{*2}}\right)\sqrt{s^2+\omega_s^2}\right]e^{-w_o}}{2\prod_{\mu=1}^{l}\sqrt{u_\mu^2-s^2}\prod_{i=1}^{g}\sqrt{(s_i^2-s^2)(s_i^{*2}-s^2)}}, \tag{28}$$

where the upper signs hold if the γ's are positive and the lower signs if the γ's are negative. The w_o equals zero if $\sum_{\mu=1}^{l}\gamma_\mu$ is odd--that is, if the degree of E(s) is odd--and it equals $\pm j\frac{\pi}{2}$ if the degree of E(s)

is even, so $\sum_{\mu=1}^{l} |\gamma_\mu|$ is even, if there is to be an attenuation pole at infinity. It is easily checked, by letting l=0, g=1, $w_o=j\frac{\pi}{2}$, that if there is an attenuation pole at ∞, there is a branch point of order two at ω_s. This is not permissible, so w_o must be zero in all cases, although this will result in finite attenuation, α_m, at infinity if the degree, N, of E(s) is even. If N is odd, the attenuation at infinity will be infinite. Another form of $[H(s)]^{-1}$ follows from using (25), with all γ_μ and γ_i equal to ½, for comparison with other works:

$$[H(s)]^{-1} = e^{-\alpha_m}\cosh\left[\sum_{\mu=1}^{l} \tfrac{1}{2}\cosh^{-1}\left(\frac{u_\mu^2+2\omega_s^2+s^2}{u_\mu^2-s^2}\right) + \sum_{i=1}^{g} \cosh^{-1}\left(\frac{\sqrt{(\omega_s^2+s_i^2)(\omega_s^2+s_i^{*2})}+\omega_s^2+s^2}{\sqrt{(s_i^2-s^2)(s_i^{*2}-s^2)}}\right)\right]. \quad (29)$$

With $s_i=u_i+jv_i$, and use of H(0)=1, one obtains

$$\alpha_m = \ln\cosh\left[\sum_{\mu=1}^{l} \tfrac{1}{2}\cosh^{-1}\left(\frac{u_\mu^2+2\omega_s^2}{u_\mu^2}\right) + \sum_{i=1}^{g} \cosh^{-1}\left(\frac{\sqrt{(\omega_s^2+u_i^2-v_i^2)+4u_i^2v_i^2}+\omega_s^2}{\sqrt{(u_i^2-v_i^2)^2+4u_i^2v_i^2}}\right)\right]. \quad (30)$$

If l=1, $s=j\omega$, and $\omega_s \leq \omega \leq \infty$,

$$[S(\omega^2)]^{-1} = e^{-\alpha_m}\cos\left[\tfrac{1}{2}\cos^{-1}\left(\frac{u_1^2+2\omega_s^2-\omega^2}{u_1^2+\omega^2}\right) + \sum_{i=1}^{g} \cos^{-1}\left(\frac{\sqrt{(\omega_s^2+u_i^2-v_i^2)^2+4u_i^2v_i^2}+\omega_s^2-\omega^2}{\sqrt{(\omega^2+u_i^2-v_i^2)^2+4u_i^2v_i^2}}\right)\right]. \quad (31)$$

Equation (30), with l=1, and equation (31), are identical to Trifinov's (17) and (16), respectively.[7] For $0 \leq \omega \leq \omega_s$, one replaces "cos" by "cosh" in (31) to obtain Trifinov's (18). Equations (30), with l=0, and equation (31) with the first term in brackets omitted, are identical to Beletskiy's (14) and (16), respectively.[6] It should be noted, however, that Beletskiy's solution, for N even, and no real natural frequencies, resulted in attenuation α_m at $s=\infty$. Unbehauen[2] has given the only known solution for N even, although he must solve nonlinear equations iteratively. For N odd equation (28) (with $w_o=0$), with $\sqrt{E(s)E(-s)}$ in the denominator replaced by E(s), and α_m given by (30), is a closed-form solution for the (reciprocal of) the operating voltage ratio or transducer function. This is a more convenient solution, for N odd, than that given by Unbehauen[2] for nonlinear equations must be solved when using his method. For N even Unbehauen's remains the only solution yielding infinite loss at infinity. However, it is hoped that this section points the way to a closed-form solution for N even as well.

To obtain Chebyshev pass-band attenuation when E(s) is prescribed,[5] one sets up

$$\frac{dw}{ds} = \left\{ \sum_{\mu=1}^{l} \frac{2\gamma_\mu \sqrt{(u_\mu^2+f_2^2)(u_\mu^2+f_1^2)}}{u_\mu^2 - s^2} \right.$$

$$\left. + \sum_{i=1}^{g} 2\gamma_i \left[\frac{\sqrt{(s_i^2+f_2^2)(s_i^2+f_1^2)}}{s_i^2 - s^2} + \frac{\sqrt{(s_i^{*2}+f_2^2)(s_i^{*2}+f_1^2)}}{s_i^{*2} - s^2} \right] \right\} \frac{s}{\sqrt{(s^2+f_2^2)(s^2+f_1^2)}} . \quad (32)$$

Integration and manipulation yields

$$w(s) = \sum_{\mu=1}^{l} 2|\gamma_\mu| \ln \left[\frac{\sqrt{(s^2+f_2^2)(u_\mu^2+f_1^2)} + \frac{\gamma_\mu}{|\gamma_\mu|} \sqrt{(s^2+f_1^2)(u_\mu^2+f_2^2)}}{\sqrt{(f_2^2-f_1^2)(u_\mu^2-s^2)}} \right]$$

$$+ \sum_{i=1}^{g} 2|\gamma_i| \ln \left\{ \left[\frac{\sqrt{(s^2+f_2^2)(s_i^2+f_1^2)} + \frac{\gamma_i}{|\gamma_i|} \sqrt{(s^2+f_1^2)(s_i^2+f_2^2)}}{\sqrt{(f_2^2-f_1^2)(s_i^2-s^2)}} \right] \right.$$

$$\left. \times \left[\frac{\sqrt{(s^2+f_2^2)(s_i^{*2}+f_1^2)} + \frac{\gamma_i}{|\gamma_i|} \sqrt{(s^2+f_1^2)(s_i^{*2}+f_2^2)}}{\sqrt{(f_2^2-f_1^2)(s_i^{*2}-s^2)}} \right] \right\} . \quad (33)$$

An approximant with Chebyshev approximation of zero in $[f_1, f_2]$ and $[-f_2, -f_1]$ is found from

$$f(s) = \cosh[w(s)] . \quad (34)$$

As in the usual band-pass case

$$\sum_{\mu=1}^{l} |\gamma_\mu| = \text{an integer.}$$

If the zeros of E(s) are counted as if they were distinct even though they are not, one may set $|\gamma_\mu| = |\gamma_i| = \frac{1}{2}$ for all μ and i, and obtain

$$f(s) = \frac{\prod_{\mu=1}^{l} \left[\sqrt{(s^2+f_2^2)(u_\mu^2+f_1^2)} \pm \sqrt{(s^2+f_1^2)(u_\mu^2+f_2^2)} \right]}{\prod_{\mu=1}^{l+2g} (f_2^2-f_1^2) \sqrt{E(s)E(-s)}}$$

$$\times \prod_{i=1}^{g} \left\{ \left[\sqrt{(s^2+f_2^2)(s_i^2+f_1^2)} \pm \sqrt{(s^2+f_1^2)(s_i^2+f_2^2)} \right] \right.$$

$$\left. \times \left[\sqrt{(s^2+f_2^2)(s_i^{*2}+f_1^2)} \pm \sqrt{(s^2+f_1^2)(s_i^{*2}+f_2^2)} \right] \right\}$$

$$+ \prod_{\mu=1}^{l} \left[\sqrt{(s^2+f_2^2)(u_\mu^2+f_1^2)} \mp \sqrt{(s^2+f_1^2)(u_\mu^2+f_2^2)} \right]$$

$$\times \prod_{i=1}^{g} \left\{ \left[\sqrt{(s^2+f_2^2)(s_i^2+f_1^2)} \mp \sqrt{(s^2+f_1^2)(s_i^2+f_2^2)} \right] \right.$$

$$\left. \times \left[\sqrt{(s^2+f_2^2)(s_i^{*2}+f_1^2)} \mp \sqrt{(s^2+f_1^2)(s_i^{*2}+f_2^2)} \right] \right\} , \quad (35)$$

where the upper signs hold for γ's equal to $\frac{1}{2}$ and the lower signs for γ's equal to $-\frac{1}{2}$. Equation (35) shows that f(s) must have even degree polynomials in numerator and denominator. One can, therefore, replace

$E(s)E(-s)$ by $[E(s)]^2$ in (35) to obtain a function $h(s)$. A suitable (reciprocal of) transducer function, $[H(s)]^{-1}$, yielding equal-ripple attenuation in the pass band may then be obtained by using[5]

$$S(s) = h(s) - B = \frac{M(s)}{E(s)} , \tag{36}$$

where $B=\frac{r}{r-1}$, and r is the ratio of the maximum to minimum magnitude of $[S(s)]^{-1}$ in the pass band, and then multiplying the right side of (36) by K to obtain

$$[H(s)]^{-1} = K \frac{M(s)}{E(s)}, \tag{37}$$

such that $|H(j\omega)| \geq 1$ for all ω. To obtain the low-pass case one sets $f_1=0$, $w_o=0$ if N is even and $w_o=\pm j\pi/2$ if N is odd. Furthermore, if N is odd, one finds[5]

$$S(s) = f(s)f(-s) - B = \frac{M(s)}{E(s)E(-s)}, \tag{38}$$

and

$$[H(s)]^{-1} = K \frac{M(s)}{[E(s)]^2} . \tag{39}$$

Considerable simplification occurs by using the z-plane given by (1). Then, with all $|\gamma|$'s equal to ½,

$$w(z) = \sum_{j=1}^{N} \ln\left(\frac{z + \frac{\gamma_j}{|\gamma_j|} z_j}{z - \frac{\gamma_j}{|\gamma_j|} z_j} \right) + w_o, \tag{40}$$

so, using $f(z)=\cosh[w(z)]$,

$$f(z) = \frac{\mathrm{Ev}\left[\prod_{j=1}^{n} \left(z + \frac{\gamma_j}{|\gamma_j|} z_j \right) \right]}{\prod_{j=1}^{n} \sqrt{(z^2 - z_j^2)}}, \tag{41}$$

for N even and $w_o=0$, and

$$f(z) = j \frac{\mathrm{Ev}\left[\prod_{j=1}^{N} \left(z + \frac{\gamma_j}{|\gamma_j|} z_j \right) \right]}{\prod_{j=1}^{N} \sqrt{(z^2 - z_j^2)}} , \tag{42}$$

for N odd and $w_o=j\pi/2$.

For the band-pass case only (41) is possible, and a more explicit form is

$$f(z)=h(z)= \frac{\mathrm{Ev}\left\{ \prod_{\mu=1}^{l} \left[z + \frac{\gamma_\mu}{|\gamma_\mu|} z_\mu \right]^{2|\gamma_\mu|} \cdot \prod_{i=1}^{g} \left[\left(z + \frac{\gamma_i}{|\gamma_i|} z_i \right)\left(z + \frac{\gamma_i}{|\gamma_i|} z_i^* \right) \right]^{2|\gamma_i|} \right\}}{\prod_{\mu=1}^{l} (z - z_\mu)^{|\gamma_\mu|} \prod_{i=1}^{g} \left[(z - z_i)(z - z_i^*) \right]^{|\gamma_i|}} . \tag{43}$$

If one attempts to obtain Chebyshev stop-band attenuation in the band-pass case, when E(s) is prescribed, one obtains a double pass band problem explicitly given by Watanabe.[3] For other multiband problems one obtains similar formulations involving elliptic or hyper-elliptic integrals of the first and second kinds for which no closed-form solution is possible, in general.

ACKNOWLEDGMENTS

Much of this work appears in the writer's Ph. D. thesis, which was accepted by the Dept. of Electrical Engineering at the University of Wisconsin, Madison, Wisconsin, May, 1968. The writer thanks Prof. T.J. Higgins for his encouragement, and the Ford Foundation and the Engineering Experiment Station for financial assistance. Further work was done since the writer has been in the Dept. of Electrical Engineering at the University of Toronto.

REFERENCES

[1] H.J. Orchard and G.C. Temes, "Filter design using transformed variables," IEEE Trans. on Cct. Theory, vol. CT-15, pp. 385-408, Dec. 1968.

[2] R. Unbehauen, "Low-pass filters with predetermined phase or delay and Chebyshev stopband attenuation," IEEE Trans. on Cct. Theory, vol. CT-15, pp. 337-341, Dec. 1968.

[3] H. Watanabe, "Approximation theory for filter networks," IEEE Trans. on Cct. Theory, vol. CT-9, pp. 341-356, Sept. 1961.

[4] M. Sablatash, "Determination of characteristic functions for crystal-capacitor band-pass ladder filters with nearly Chebyshev pass-band attenuation," Proc. 1964 2nd Allerton Conf. on Cct. and System Theory, pp. 149-174.

[5] B.J. Bennett, "Synthesis of electric filters with arbitrary phase characteristics," IRE Nat. Conv. Rec., pt. 5, Cct. Theory, pp. 19-26, 1953.

[6] A.F. Beletskiy, "Synthesis of filters with linear phase characteristics," Telecommunications and Radio Engineering, vol. 15, pt. 1, pp. 39-48, April 1961.

[7] I.I. Trifinov, "Synthesis of low-frequency filters with assigned phase characteristics, "Telecommunications and Radio Engineering, vol. 16, pt. 1, pp. 60-68, Jan. 1962.

ANALYSIS AND SYNTHESIS OF FILTERS BY TRANSFORMATION OF MATRIX BINOMIALS

THOMAS G. MARSHALL, JR.
Rutgers University
New Brunswick, New Jersey

ABSTRACT

A matrix binomial, $s\underline{L} - \underline{K}$, is introduced which corresponds to an RLC lowpass filter with finite transmission zeros. An equivalence transformation yields a tridiagonal matrix which is basic to the algebraic analysis and synthesis procedures that are developed. The procedures apply to lossy filters and do not require complex arithmetic or the formation of polynomials in s.

An algorithm is presented for finding polynomials x and y in the classical algebraic equation $a = xb + yc$. This algorithm and a vector space analog replace the zero-shifting procedure for realizing transmission zeros in the synthesis procedures that are described.

I. INTRODUCTION

Algebraic synthesis procedures have previously been discussed which apply to RLC lowpass filters without finite transmission zeros[1]. In this paper, this work is extended to the analysis and synthesis of RLC lowpass filters having finite transmission zeros. Losses can be associated with zero forming branches, so complex zeros of transmission can be considered.

The algebraic problems which are encountered are quite interesting. The matrix $s\underline{L} - \underline{K}$ bears a relation to a tridiagonal matrix, so the synthesis procedure is a generalization of the recursive procedures associated with tridiagonalization and the Euclidean algorithm. The algorithm referred to in the abstract plays an analogous role in the synthesis procedure to that of the division algorithm in the Euclidean algorithm [2].

II. A PAIR OF MATRIX BINOMIALS FOR A LADDER NETWORK

Lowpass filters with n branches, some of which might be zero forming branches, will be considered. Network N in Figure 1 is an example in which ρ_1, ρ_{13}, etc. are conductance values, ρ_2 and ρ_4 are resistance values, and ϵ_i and $\epsilon_{i-1,i+1}$ are either capacitance or inductance values, as indicated. Certain elements such as ϵ_i or ρ_{13} or ϵ_{13} can be zero subject to

the restriction that the order of the network be n.

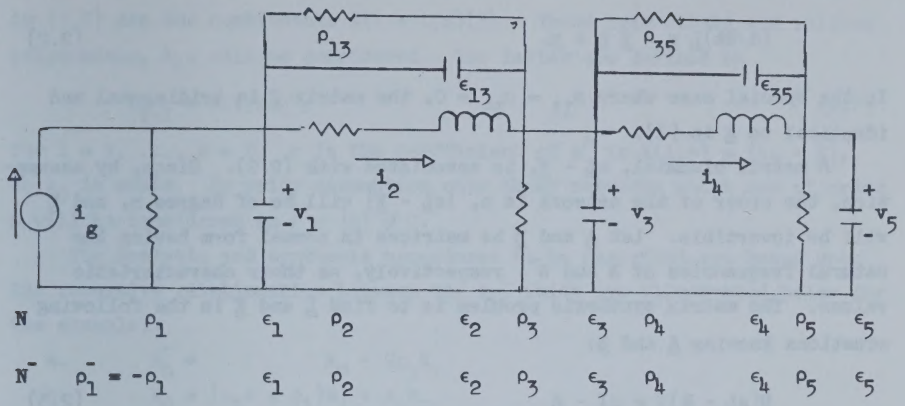


Figure 1. A Typical Network to be Considered

The network $\bar{N}$ in Figure 1 is introduced to facilitate the formulation of the matrix synthesis problem. It can be seen to have the zeros of reflection of N for its natural frequencies [1], so the problem of realizing a specified reflection coefficient, s_{11}, with N can be restated more symmetrically as follows: Synthesize the pair of networks N and $\bar{N}$ which realize, respectively, the poles and zeros of s_{11} as their natural frequencies and which also realize specified transmission zeros.

If Kirchhoff's voltage and current laws are employed alternately [3], the following matrix differential equation is obtained for network N:

$$\frac{d}{dt}\begin{bmatrix} (\epsilon_1+\epsilon_{13}) & -\epsilon_{13} & & & \\ -\epsilon_{13} & \epsilon_2+\epsilon_3+\epsilon_{13}+\epsilon_{35} & -\epsilon_{35} & & \\ & -\epsilon_{35} & \epsilon_4+\epsilon_5+\epsilon_{35} & & \end{bmatrix}\begin{bmatrix} v_1 \\ i_2 \\ v_3 \\ i_4 \\ v_5 \end{bmatrix}$$

$$= \begin{bmatrix} (-\rho_1-\rho_{13}) & -1 & \rho_{13} & & \\ 1 & -\rho_2 & -1 & & \\ \rho_{13} & 1 & (-\rho_3-\rho_{13}-\rho_{35}) & -1 & \rho_{35} \\ & & 1 & -\rho_4 & -1 \\ & & \rho_{35} & 1 & (-\rho_5-\rho_{35}) \end{bmatrix}\begin{bmatrix} v_1 \\ i_2 \\ v_3 \\ i_4 \\ v_5 \end{bmatrix} + \begin{bmatrix} i_g \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (2.1)$$

The matrices and column vectors in (2.1) will be denoted as shown in (2.2):

$$(d/dt)\underline{L}\,\underline{x} = \underline{K}\,\underline{x} + \underline{u}. \tag{2.2}$$

In the special case where $\rho_{13} = \rho_{35} = 0$, the matrix $\underline{K}$ is tridiagonal and identical to $\underline{K}$ in [1].

A matrix binomial, $s\underline{L} - \underline{K}$, is associated with (2.2). Since, by assumption, the order of the network is n, $|s\underline{L} - \underline{K}|$ will be of degree n, and $\underline{L}$ will be invertible. Let $\underline{A}$ and $\underline{B}$ be matrices in normal form having the natural frequencies of N and $\bar{N}$, respectively, as their characteristic values. The matrix synthesis problem is to find $\underline{L}$ and $\underline{K}$ in the following equations knowing $\underline{A}$ and $\underline{B}$:

$$\underline{U}(s\underline{L} - \underline{K})\underline{V} = s\underline{I} - \underline{A} \tag{2.3}$$

and

$$\underline{\bar{U}}(s\underline{L} - \underline{\bar{K}})\underline{\bar{V}} = s\underline{I} - \underline{B}, \tag{2.4}$$

where superscript "-" corresponds to $\bar{N}$. Thus $\underline{\bar{K}}$ differs from $\underline{K}$ only in the sign of entry (1,1). Since s appears only where explicitly shown in the binomials, the matrices $\underline{U}$, $\underline{V}$, $\underline{\bar{U}}$, and $\underline{\bar{V}}$ can all be independent of s [4]. Subject to slight restrictions, to be discussed in Section IV, $\underline{A}$ and $\underline{B}$ will be cyclic (nonderogatory) matrices.

A tridiagonal matrix of degree 2 in s, equivalent to $s\underline{L} - \underline{K}$, can be obtained by postmultiplying this matrix by an elementary matrix, $\underline{E}_0$. For the example,

$$(s\underline{L} - \underline{K})\underline{E}_0 = \begin{bmatrix} \epsilon_1 s + \rho_1 & 1 & & & \\ -z_2 & \epsilon_2 s + \rho_2 & z_2 & & \\ & -1 & \epsilon_3 s + \rho_3 & 1 & \\ & & -z_4 & \epsilon_4 s + \rho_4 & z_4 \\ & & & -1 & \epsilon_5 s + \rho_5 \end{bmatrix}, \tag{2.5}$$

where

$$z_i = 1 + (\epsilon_i s + \rho_i)(\epsilon_{i-1,i+1} s + \rho_{i-1,i+1}). \tag{2.6}$$

The matrix $\underline{E}_0$ is that which, for each zero forming branch consisting of $\rho_{i-1,i+1}$ and/or $\epsilon_{i-1,i+1}$, subtracts the product of $(\epsilon_{i-1,i+1} s + \rho_{i-1,i+1})$ times column i from column i-1 and adds the product to column i+1. It can be seen that, in general, $(s\underline{L} - \underline{K})\underline{E}_0$ will then be a tridiagonal matrix, as illustrated in (2.5), with z_i as given by (2.6). It is interesting to compare (2.5) with the tridiagonal determinant in [3].

The principal minors formed by deleting the first i rows and columns in (2.5) are the continuants K(i + 1,n)[5]. These polynomials and related polynomials, k_i, will be considered. The latter are defined by

$$k_{i-1} = K(i,n)/\gamma \tag{2.7}$$

for i = 1, ..., n + 2. γ is the coefficient of s^n in $K(1,n) = |s\underline{L} - \underline{K}|$, so k_0 is monic. By prior assumption only those networks which are of order n will be considered so $\gamma = |\underline{L}| \neq 0$.

The analysis and synthesis procedures to be described are based upon the recursive relationships between the k_i, which are illustrated below for the example:

$$\begin{aligned}
&\text{a.} \quad \bar{k}_0 = k_0 - 2\rho_1 k_1 \\
&\text{b.} \quad k_0 = (\epsilon_1 s + \rho_1)k_1 + z_2 k_2 \\
&\text{c.} \quad k_1 = (\epsilon_2 s + \rho_2)k_2 + z_2 k_3 \\
&\text{d.} \quad k_2 = (\epsilon_3 s + \rho_3)k_3 + z_4 k_4 \\
&\text{e.} \quad k_3 = (\epsilon_4 s + \rho_4)k_4 + z_4 k_5 \\
&\text{f.} \quad k_4 = (\epsilon_5 s + \rho_5)k_5 \\
&\text{g.} \quad k_5 = 1/\gamma
\end{aligned} \tag{2.8}$$

All but the first of these equations is obtained by expanding minor K(i + 1,n) along its top row. The polynomial $\bar{k}_0$ is defined analogously to k_0 so that

$$\bar{k}_0 = |s\underline{L} - \underline{\bar{K}}|/\gamma = (\epsilon_1 s - \rho_1)k_1 + z_2 k_2 , \tag{2.9}$$

from which the first equation in (2.8) is obtained. In general,

$$k_{i-1} = (\epsilon_i s + \rho_i)k_i + z k_{i+1} , \tag{2.10}$$

where z is either z_i, z_{i+1}, or unity. The recursive relationship between k_{i-1}, k_i, and k_{i+1}, forms the basis for both the analysis and synthesis procedures to follow.

III. ANALYSIS

A method will be described for determining the remainder, r_0, in

$$k_0 = q_0 m_0 + r_0, \tag{3.1}$$

directly from the ϵ and ρ values without first finding k_0. An example illustrates applications to finding the factors of k_0, finding the coefficients of k_0, and evaluating k_0 at a complex frequency, λ. In order to avoid complex arithmetic in finding r_0 in (3.1) for the latter problem, the divisor with real coefficients,

$$m_0 = (s - \lambda)(s - \lambda^*), \tag{3.2}$$

is employed instead of $s - \lambda$.

The remainder r_0 will be determined, recursively, by a generalization of the procedure used for finding the remainder of the characteristic polynomial of a Hessenberg matrix [6]. The row vectors

$$\underline{k}_i = \underline{v}\, k_i(\underline{M}), \tag{3.3}$$

are employed, where $\underline{M}$ is a cyclic matrix with characteristic polynomial m_0 and $\underline{v}$ is a cyclic vector for $\underline{M}$. These vectors can be determined recursively from the relation

$$\underline{k}_{i-1} = \underline{k}_i(\epsilon_i \underline{M} + \rho_i \underline{I}) + \underline{k}_{i+1} z(\underline{M}), \tag{3.4}$$

which follows directly from (2.10). The following example illustrates the procedure for both k_0 and $\bar{k}_0$:

Example 1: A singly-terminated filter will be considered with $\rho_1 = 1$, $\rho_2 = \rho_3 = \rho_{13} = 0$, $\epsilon_1 = 2/7$, $\epsilon_2 = 49/48$, $\epsilon_3 = 12/7$, and $\epsilon_{13} = 12/49$. Comparison with (2.1) indicates that

$$s\underline{L} - \underline{K} = \begin{bmatrix} 26/49 & & -12/49 \\ & 49/48 & \\ -12/49 & & 96/49 \end{bmatrix} - \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \tag{3.5}$$

and, from (2.6),

$$z_2 = 1 + s^2/4 \ . \tag{3.6}$$

It can be seen that $\gamma = |L| = 1$.

Two choices of $\underline{M}$ and $\underline{v}$ will be considered. They are

$$\underline{M} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \left[\begin{array}{cc:cc} & & \multicolumn{2}{c}{\multirow{2}{*}{$\cdots\cdots$}} \\ & \multicolumn{1}{c:}{\multirow{2}{*}{0}} & & \\ \hdashline & & 0 & -1 \\ & & 1 & 0 \end{array}\right], \tag{3.7}$$

and

$$\underline{v} = [0\ 0\ 1],\ [1 \vdots 0\ 1]. \tag{3.8}$$

For the first choice of $\underline{M}$ and $\underline{v}$, $m_0 = s^3$, and the coefficients of the k_i will be the entries of $\underline{k}_i$. In the second choice $m_0 = s(s^2 + 1)$. The factors s and $(s^2 + 1)$ are simultaneously considered by juxtaposing the two matrices and vectors, which are separated by dotted lines in (3.7) and (3.8), into one.* The coefficients of both remainders are accordingly juxtaposed in the second vector $\underline{k}_i$ which follows.

Equation (3.4) yields,

$$\underline{k}_3 = (1/r)\underline{v} \qquad = [0\ 0\ 1],\ [1\vdots\ 0\ \ 1]$$

$$\underline{k}_2 = \underline{k}_3(\epsilon_3\underline{M}) \qquad = [0\ 12/7\ 0],\ [0\vdots\ 12/7\ \ 0]$$

$$\underline{k}_1 = \underline{k}_2(\epsilon_2\underline{M}) + \underline{k}_3(\underline{I} + \underline{M}^2/4) = [2\ 0\ 1],\ [1\vdots\ 0\ -1]$$

$$\underline{k}_0 = \underline{k}_1(\epsilon_1\underline{M} + \rho_1\underline{I}) + \underline{k}_2(\underline{I} + \underline{M}^2/4) = [2\ 2\ 1],\ [1\vdots\ 1\ -1]$$

$$\underline{k}_0^- = \underline{k}_0 - 2\rho_1\underline{k}_1 \qquad = [-2\vdots\ 2\ -1],\ [-1\vdots\ 1\ \ 1]$$

From the first vector, $\underline{k}_0$, it is seen that k_0 is the Butterworth polynomial of degree 3, $s^3 + 2s^2 + 2s + 1$. From the second vector, $\underline{k}_0$, it is seen that the remainders of k_0/s and $k_0/(s^2 + 1)$ are respectively 1 and s-1. These remainders yield the values $k_0(0) = 1$ and $k_0(j1) = j - 1$.

The three individual matrices considered here are companion matrices of the factors s^n, s, and $s^2 + 1$, and the corresponding vectors are respectively the unit row vectors $\underline{e}_3$, $\underline{e}_1$ and $\underline{e}_2$. Other choices of $\underline{M}$ and $\underline{v}$ for these same factors could be selected to change the representation of the remainders, r_i[1,7].

IV. AN ALGORITHM FOR SOLVING a = xb + yc

A. Applications of the Algorithm to Synthesis

The synthesis procedure for the class of filters considered here is similar to a synthesis procedure for the class without finite transmission zeros, except that a sequence of polynomials, such as (2.8), replaces the sequence generated by the Euclidean algorithm [1]. The discussion will be limited to solving those equations (2.10) for which $z \neq 1$ which satisfy additional restrictions to be stated, the other equations being solvable by the division algorithm or reducible to the case at hand.

The equations (2.10) to be considered are of the form,

$$a = xb + yc, \tag{4.1}$$

where a, b, and c are known, nonzero, polynomials, and x and y are unknown polynomials to be determined. It is assumed that deg (a) < deg (b) + deg(c) and that b and c (which correspond respectively to k_i and z in (2.10)) are relatively prime. The latter assumption is also necessary in the zero-shifting procedure. With these assumptions it is known that unique solutions x and y exist with deg(x) < deg(c) and deg(y) < deg(b). The related problem of expressing a/bc in the form

$$a/bc = x/c + y/b, \tag{4.2}$$

---

* Since a complete set of trial factors can be simultaneously considered, this analysis procedure has application in global procedures. It can also be used for Hessenberg matrices.

is also solved, by solving (4.1). A common method of solution is to equate numerator coefficients in (4.2), after placing the right hand member on the denominator, bc.

It can be shown** that a sufficient condition for matrices $\underline{A}$ and $\underline{B}$ to be cyclic, as assumed in Section II, is for each zero forming factor, z_i, to be relatively prime to k_{i-1}. Comparison with the assumptions of this section reveals for any ladder network of the class considered here that can be synthesized by the procedure of this paper or the zero-shifting method that the matrices $\underline{A}$ and $\underline{B}$ in (2.3) and (2.4) are necessarily cyclic. This further implies that system matrices for these networks such as $\underline{L}^{-1}\underline{K}$, $\underline{K}\,\underline{L}^{-1}$, or Bashkow's A matrix are cyclic.

B. Basis for the Algorithm

When a = 1, Euclid's algorithm can be used to find x and y [8]. This procedure can be generalized by observing that there are two cases in general. After possible relabeling, it may be assumed that deg(c) $\leq$ deg(b).

Case I: c divides a (as when deg(c) = 0).

Then

$$x = 0, \tag{4.4}$$

and

$$y = a/c. \tag{4.5}$$

With the convention that deg(0) = $-\infty$, the conditions deg(x) < deg(c) and deg(y) < deg(c) are satisfied, so the solution (4.4) and (4.5) is the desired, unique solution.

In the alternative case, a and b are divided by c, giving

$$a = q_0 c + r_0 \tag{4.6}$$

and

$$b = q_1 c + r_1 \tag{4.7}$$

which permits the statement of

Case II: c does not divide a.

Then, the polynomials in

$$r_0 = x'c + y'r_1 = x'b' + y'c' \tag{4.8}$$

satisfy the same assumptions as do those in (4.1), and x and y can be obtained from

$$x = y' \tag{4.9}$$

** The author's proof of this is based upon a development in terms of elementary transformations. This development is not otherwise relevant to this paper and will be omitted.

and

$$y = q_0 - y'q_1 + x'. \tag{4.10}$$

Since b and c are relatively prime, $c' = r_1 \neq 0$, and $\deg(c') < \deg(c)$. If Case I applies to (4.8), the unknowns x' and y' can be obtained; otherwise, further repetitions of the procedure are made until Case I is obtained. The solutions for each step are then substituted successively in (4.9) and (4.10), until the original x and y are obtained.

The justification for this procedure is obtained by solving for xb in (4.1) and using (4.6), and comparing with x times (4.7). In the latter equation, the term xr_1 is divided by c giving

$$xr_1 = q_2c + r_2. \tag{4.11}$$

Comparison with the former equation indicates that $r_2 = r_0$. Making the association $x' = -q_2$, it is seen that (4.11) and (4.8) are identical and that (4.9) and (4.10) are implied.

C. The Algorithm

The preceding discussion applies to any a, b, and c satisfying the stated conditions. The algorithm will be stated for the specific case of most interest in filter synthesis in which c is quadratic. To simplify the notation, the algorithm will be stated for the particular equation (2.10),

$$k_0 = (\epsilon_1 s + \rho_1)k_1 + k_2 z_2. \tag{4.12}$$

In this case q_2 is a constant, denoted α. It is seen from (2.8) that two repetitions of this algorithm are required for each zero forming factor which is not unity.

Algorithm 1.: Given k_0, k_1, and z_2, find $(\epsilon_1 s + \rho_1)$ and k_2.

(1) Find q_0, q_1, r_0 and r_1.

(2) a. If $r_1 = 0$, stop (hypothesis not satisfied).
b. If $r_0 = 0$, let $\alpha = \epsilon_1 = \rho_1 = 0$, and go to 4.
c. If $\deg(r_1) = 0$, let $\alpha = 0$ and $\epsilon_1 s + \rho_1 = r_0/r_1$, and go to 4.
d. If $\deg(r_1) = 1$, form

$$r_0 = q_0'r_1 + r_0'$$
$$z_2 = q_1'r_1 + r_1'$$
$$\alpha = -r_0'/r_1'.$$

(3) $\epsilon_1 s + \rho_1 = q_0' + \alpha q_1'$

(4) $k_2 = q_0 - (\epsilon_1 s + \rho_1)q_1 - \alpha$

The following example will illustrate the two required applications of

Algorithm 1 corresponding to a single zero forming factor. The polynomials of Example 1 will be used.

Example 2:

Use Algorithm 2.

(1) $k_0 = (4s + 8)z_2 + (-2s - 7)$

$k_1 = 8z_2 + -7$

(2c) $\alpha = 0$, $\epsilon_1 s + \rho_1 = (-2s - 7)/(-7) = (2/7)s + 1$

(4) $k_2 = (4s + 8) - (\epsilon_1 s + \rho_1)8 = (12/7)s$

Use Algorithm 2.

(1) $k_1 = 8z_2 + -7$

$k_2 = 0z_2 + (12/7)s$

(2d) $r_1 = 0[(12/7)s] - 7$

$z_2 = [(7/48)s][(12/7)s] + 1$

$\alpha = -(-7/1) = 7$

(3) $\epsilon_2 s + \rho_2 = 0 + 7(7/48)s$

(4) $k_3 = 8 - 0 - 7 = 1$

V. TRANSFORMATION SYNTHESIS

The transformations of (2.3) and (2.4) are implicit in the synthesis procedure to be described. This procedure has two steps which effectively decompose $\underline{U}$, $\underline{V}$, $\underline{U}^-$, and $\underline{V}^-$ so that (2.3) and (2.4) become,

$$\underline{R}\ \underline{P}(s\underline{L} - \underline{K})\underline{Q}\ \underline{R}^{-1} = s\underline{I} - \underline{A} \tag{5.1}$$

and

$$\underline{S}\ \underline{P}(s\underline{L} - \underline{K}^-)\underline{Q}\ \underline{S}^{-1} = s\underline{I} - \underline{B}. \tag{5.2}$$

In the first step of the procedure, the matrices $\underline{R}$ and $\underline{S}$ are determined as described elsewhere [1]. The vectors $\underline{k}_i$ which are determined in the second step are closely related* to the rows $\underline{q}_i$ of $\underline{Q}$. The matrices $\underline{R}^{-1}$, $\underline{S}^{-1}$, and $\underline{P}$ do not appear. The synthesis procedure, itself, will be described below; however, the derivation of (5.1) and (5.2) will be omitted.

The vectors $\underline{k}_i$ are related in the synthesis procedure, as in the analysis procedure, to the polynomials k_i through (3.3). It will now be

---

* Where there are no finite transmission zeros they are equal. See equations (59) and (60) in [1].

assumed that $\underline{M}$ is an n x n unit-upper-Hessenberg (U.U.H.) matrix and that $\underline{v}$ is the unit row vector $\underline{e}_n$. $\underline{A}$ and $\underline{B}$ are likewise assumed to be U.U.H. matrices, which is consistent with their cyclic property, so that $\underline{M} = \underline{A}$ or $\underline{M} = \underline{B}$ are possible choices. In these special cases, $\underline{k}_0 = 0$ or $\underline{k}_0^- = 0$ and $\underline{R} = \underline{I}$ or $\underline{S} = \underline{I}$, respectively.

Step 1 of Algorithm 1 requires the solution of equations of the form

$$\underline{k}_{i-2} = \underline{q}_i z(\underline{M}) + \underline{r}_{n-1}, \tag{5.3}$$

where the subscripts now correspond to the degree of the polynomial associated with the vector, e.g. $\deg(q_i) = n - i$. This equation will have a unique solution subject to the stipulation that the vectors in (5.3) be those unique vectors determined by the polynomials in the equation

$$k_{i-2} = q_i z + r_{n-1}. \tag{5.4}$$

The assumption that $\underline{M}$ be U.U.H. permits (5.3) to be solved analogously to polynomial division by determining a sequence of partial remainders,

$$\underline{r}_{j-1} = \underline{r}_{j-2} - q_{ij}\underline{e}_j z(\underline{M}), \tag{5.5}$$

for $j = i, \ldots, n$, in which $\underline{r}_{i-2} = \underline{k}_{i-2}$ initiates the procedure. In general the entry q_{ij} of $\underline{q}_i$ is the value that causes entry $r_{j-1,j-2}$ of $\underline{r}_{j-1}$ to be zero. In the special case where $i = 2$, $q_{22} = 1/(\epsilon_2\epsilon_{13})$, which follows from (2.6) and (5.4).

The polynomials r_0 or r_1 are easily determined from the vectors $\underline{r}_{n-1}$ in (5.3) [1,7], so steps 2 and 3 of Algorithm 1 needn't be altered. Step 4 merely involves the substitution of vectors for polynomials and $\underline{M}$ for s. The following example illustrates these points and the complete synthesis procedure.

Example 3: Synthesize the network of Example 1. Let $\underline{M} = \underline{A}$ be the tridiagonal normal form of Example 2 in [1], and let $\underline{B}$ be determined from $\underline{A}$ by changing the signs of the main diagonal entries. Then their characteristic polynomials will be k_0 and k_0^- of Example 1, as desired.

Step 1. From Algorithm 1 in [1], $\underline{k}_0 = 0$ and $\underline{k}_0^- = [-4\ 4\ -6]$.

Step 2. Determine element values and $\underline{k}_1$. Specify $\rho_1 = 1$. (Steps b and c are Algorithm 1. Compare with Example 2.).

a. $\underline{k}_1 = [2\ -2\ 3]$, from (2.8a).

b. $\underline{k}_0 = [0\ 4\ 4](\underline{I} + \underline{M}^2/4) + [0\ -2\ -5]$.

$\underline{k}_1 = [0\ 0\ 8](\underline{I} + \underline{M}^2/4) + [0\ 0\ -7]$.

$\epsilon_1 = 2/7$. Same calculations as Example 2.

$\underline{k}_2 = [0\ 4\ 4] - [0\ 0\ 8](2/7\ \underline{M} + \underline{I}) - 0\underline{e}_3 = [0\ 12/7\ -12/7]$.

c. $\underline{k}_1 = [0\ 0\ 8](\underline{I} + \underline{M}^2/4) + [0\ 0\ -7]$.

$\underline{k}_2 = [0\ 0\ 0](\underline{I} + \underline{M}^2/4) + [0\ 12/7\ -12/7]$.

$\epsilon_2 = 49/48$, $\rho_2 = 0$. Same calculations as Example 2.

$\underline{k}_3 = [0\ 0\ 8] - [0\ 0\ 0](49/48\ \underline{M}) - 7\underline{e}_3 = [0\ 0\ 1]$.

d. $\epsilon_{13} = 12/49$, $\rho_{13} = 0$, from (2.6).

e. $\epsilon_3 = 12/7$, $\rho_3 = 0$ from (2.8d).

VI. BIBLIOGRAPHY

[1] T. G. Marshall, Jr., "Synthesis of RLC Ladder Networks by Matrix Tridiagonalization," IEEE Trans. Circuit Theory, vol. CT-16, pp.39-46, February 1969.

[2] T. G. Marshall, Jr., "A Greatest Common Divisor Algorithm for a Vector Space," Proc. 1968 Allerton Conf.

[3] T. R. Bashkow, "A Note on Ladder Network Analysis," IRE Trans. Circuit Theory (Correspondence), vol. CT-8, pp. 168-169, June 1961.

[4] F. R. Gantmacher, The Theory of Matrices, vol.1, New York: Chelsea, 1959, Ch.6, p.145.

[5] T. Muir, A Treatise on the Theory of Determinants, revised ed. by W.H. Metzler. New York: Dover, 1960, Ch.13.

[6] J. H. Wilkinson, The Algebraic Eigenvalue Problem, London: Oxford University Press, 1965, Ch.7.

[7] T. G. Marshall, Jr., "Tridiagonalization and Related Matrix and Network Algorithms for a Polynomial Space," Proc. 1967 Allerton Conf., pp. 544-556.

[8] M. Böcher, Introduction to Higher Algebra, New York: MacMillan, 1907, Ch. 15, p. 193.

RICHARDS' THEOREM FOR POSITIVE REAL MATRICES OF SEVERAL VARIABLES+

AHMED M. SOLIMAN
Electrical Engineering Department
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

ABSTRACT

An extension to Richards' Theorem for positive real matrices of several variables is shown.

I. INTRODUCTION

Positive real functions and matrices of several variables are introduced concerning the problem of designing a passive network composed of lumped elements with variable parameters[1]. Positive real matrices of several variables seem to be efficiently applicable to the study of all the problems of synthesizing networks having distributed structures or lumped distributed structures[2,3].

Richards' theorem[4] seems to be the key to very useful synthesis procedures. Newcomb[5] in 1965 gave for the first time a nonreciprocal transmission line n port synthesis based upon Bayard's matrix extension of Richards' theorem for the single variable case[6].

Saito[2] in 1966 gave an extension to Richards' theorem to multivariable positive real functions, and he used the theorem to derive the realizability conditions for some cascaded commensurable and noncommensurable transmission lines.

The present paper gives an extension to Richards' theorem for positive real matrices of several variables.

II. DEFINITIONS

Definition 1[7]: If an m x m matrix $Z(P_1,P_2,\ldots,P_n)$ is rational, real for real P_i $(1 \leq i \leq n)$, and if $Z + Z_T^*$ is non-negative Hermitian in the open polydomain $D_R \subset C^n$, then Z is called a positive real matrix of n variables.

Here a superscript asterisk $(^*)$ denotes complex conjunction, and a subscript T denotes the transposition.

Definition 2: The P_1-degree of a rational n variable matrix $Z(P_1,P_2,\ldots,P_n)$ is defined as:

$$P_1 \text{ - degree of } Z(P_1,P_2,\ldots,P_n) = \deg_{P_1} [Z(P_1,P_2,\ldots,P_n)]$$

$$= \max_{P_{20},P_{30},\ldots,P_{no}} \deg_{P_1} [Z(P_1,P_{20},\ldots,P_{no})]$$

where $\deg_{P_1} [Z(P_1,P_{20},\ldots,P_{no})]$ is the McMillan[10] degree of $Z(P_1,P_{20},\ldots,P_{no})$.

+This paper is a condensation of part of the Ph.D. dissertation to be submitted to the University of Pittsburgh.

For any fixed $P_{20},P_{30},\ldots,P_{no}$, $Z(P_1,P_{20},\ldots,P_{no})$ is a matrix of rational functions in P_1 and as such has an unambiguous McMillan degree.

III. TESTING MULTIVARIABLE POSITIVE REAL MATRICES

The following theorem extended from the single variable to multivariable case by Koga, will be written in the form of a lemma.

Lemma 1[7]

The necessary and sufficient condition for an m x m matrix $Z(P_1,P_2,\ldots,P_n)$ to be positive real matrix of n variables is that

$$S = (Z + I_m)^{-1}(Z - I_m)$$

is analytic in the open polydomain $D_R \subset C^n$ real for P_i $(1 \leq i \leq n)$ real, and

$$I_m - S_T^* S \geq 0_m$$

for Re $P_i = 0$ $(1 \leq i \leq n)$.

The following lemma will also be useful in the proof of the theorem[8].

Lemma 2

A real symmetric matrix K of rank m is positive semidefinite if and only if there exists a matrix C of rank m such that $K = C_T C$.

IV. RICHARDS' THEOREM FOR MULTIVARIABLE POSITIVE REAL MATRICES

Here we state and prove a generalization of Richards' theorem for symmetric positive real matrices of several variables.

Theorem

If $Z(P_1,P_2,\ldots,P_n)$ is an m x m nonsingular, symmetric positive real matrix of n variables $P_1,P_2,\ldots,P_n$, and

$$Z(1,P_2,\ldots,P_n) = K \quad \text{a constant m x m matrix independent of } P_2,P_3,\ldots,P_n.$$

then

$$Z_R(P_1,P_2,\ldots,P_n) \text{ which is defined by}^+$$

$$Z_R(P_1,P_2,\ldots,P_n) = K[Z(P_1,P_2,\ldots,P_n) - P_1K]^{-1}[K - P_1Z(P_1,P_2,\ldots,P_n)]$$

is also positive real, and in general

+For the theorem we assume that $[Z(P_1,\ldots,P_n) - P_1K]$ is nonsingular, the singular case can be handled exactly as the single variable case[9].

$$\deg_{P_i} [Z_R] \leq \deg_{P_i} [Z] \qquad i=1,2,\ldots,n.$$

In particular if

$$Z(1,P_2,\ldots,P_n) = -Z(-1, P_2,\ldots,P_n)$$

$$\deg_{P_1} [Z_R] = \deg_{P_1} [Z] - 1$$

Proof

The m x m matrix K is symmetric and positive semidefinite, since if we let P_j = real and positive, j=2,3,...,n, then $Z(1,P_2,\ldots,P_n)$ is a positive semidefinite m x m real matrix independent of P_j, j=2,3,...,n because $Z(P_1,P_2,\ldots,P_n)$ is positive real matrix. Thus by lemma (2), K can be factored to

$$K = C_T C$$

Since $Z(P_1,\ldots,P_n)$ is nonsingular, so is K and hence C. To simplify the proof, let us normalize[9] our matrix Z, by defining

$$Z_n(P_1,P_2,\ldots,P_n) = C_T^{-1} \cdot Z(P_1,P_2,\ldots,P_n) \cdot C^{-1}$$

Therefore

$$Z_n(1,P_2,\ldots,P_n) = C_T^{-1} \cdot Z(1,P_2,\ldots,P_n) \cdot C^{-1} = C_T^{-1} K C^{-1} = I_m$$

$$Z_{R_n}(P_1,P_2,\ldots,P_n) = C_T^{-1} Z_R(P_1,\ldots,P_n) C^{-1}$$

Thus we want to show

$$Z_{R_n}(P_1,P_2,\ldots,P_n) = [Z_n(P_1,P_2,\ldots,P_n) - P_1 I_m]^{-1}[I_m - P_1 Z_n(P_2,\ldots,P_n)]$$

is m x m positive real matrix.

The matrix $Z_{R_n}(P_1,P_2,\ldots,P_n)$ is obviously rational with real coefficients when $Z_n(P_1,P_2,\ldots,P_n)$ is.

Form the matrix $S_{R_n}(P_1,P_2,\ldots,P_n)$ defined by

$$S_{R_n}(P_1,P_2,\ldots,P_n) = (Z_{R_n} + I_m)^{-1}(Z_{R_n} - I_m)$$

Thus

$$S_{R_n} = \frac{P_1 + 1}{P_1 - 1} (Z_n + I_m)^{-1} (Z_n - I_m)$$

By defining

$$S_n = (Z_n + I_m)^{-1}(Z_n - I_m)$$

Hence

$$S_{R_n} = \frac{P_1 + 1}{P_1 - 1} S_n$$

By lemma (1), S_n is analytic in the open polydomain $D_R \subset C^n$, real for P_i $(1 \leq i \leq n)$ real, and $I_m - S_T^* S \geq 0$ for Re $P_i = 0$ $(1 \leq i \leq n)$. Therefore S_{R_n} is analytic in the open polydomain $D_R \subset C^n$. Further since for $P_1 = j\omega_1$ $\left|\frac{j\omega_1 + 1}{j\omega_1 - 1}\right| = 1$. Therefore $[S_{R_n}^*]_T [S_{R_n}] = [S_n]_T^* [S_n]$ for Re $P_i = 0$ $(1 \leq i \leq n)$.

Therefore from Lemma (1), Z_{R_n} is an m x m positive real matrix for the n variables $P_1, P_2, \ldots, P_n$.

It is obvious that $\deg_{P_i} S_{R_n} \leq \deg_{P_i} S_n$, i=1,2,...,n because (P_1-1) cancels.

Thus $\deg_{P_i} [Z_R] = \deg_{P_i} [Z_{R_n}] \leq \deg_{P_i} [Z] = \deg_{P_i} [Z_n]$ i=1,2,...,n. If in particular $Z(1,P_2,\ldots,P_n) = -Z(-1,P_2,\ldots,P_n)$, then $(P_1 + 1)$ cancels and we have

$$\deg_{P_1} [Z_R] = \deg_{P_1} [Z] - 1$$

O.E.D.

The network representation of this theorem is shown in Figure 1, where the m port unit element is defined as in Newcomb[5], and has a characteristic impedance matrix $\underline{K}$. Clearly if $\overline{Z}(P_1,P_2,\ldots,P_n)$ is an m x m nonsymmetric matrix such that $\overline{Z}(1,P_2,\ldots,P_n) = \overline{K}$ a constant m x m matrix, it can be made symmetrical at $P_1 = 1$ similar to the single variable case[9]. The physical structure for transfering the nonsymmetrical matrix $\overline{Z}(P_1,P_2,\ldots,P_n)$ to a symmetrical one $Z(P_1,P_2,\ldots,P_n)$ and of normalizing the latter to $Z_n(P_1,P_2,\ldots,P_n)$ is shown in Figure 2, where $Z_G = \frac{1}{2}[\overline{K} - \overline{K}_T]$.

Figure 3 represents a balanced bridge realization of Bott-Duffin to interpret $Z_n(P_1,P_2,\ldots,P_n)$ in terms of $Z_{R_n}(P_1,P_2,\ldots,P_n)$, which is quite the same situation as the single variable case. (We have assumed that $Z_{R_n}(P_1,P_2,\ldots,P_n)$ is nonsingular).

Corollary

If $Z(P_1,P_2,\ldots,P_n)$ is an m x m nonsingular symmetric positive real matrix of n variables $P_1,P_2,\ldots,P_n$, and if a is any positive real number such that $Z(a,P_2,\ldots,P_n) = K$ a constant m x m matrix independent of $P_2,P_3,\ldots,P_n$.

Then

$$Z_R(P_1,P_2,\ldots,P_n) = K[a\,Z(P_1,P_2,\ldots,P_n) - P_1K]^{-1}[a\,K - P_1Z(P_1,P_2,\ldots,P_n)].$$

is also positive real[+], and in general

[+]Assume that $[a\,Z(P_1,P_2,\ldots,P_n) - P_1K]$ is nonsingular. Again if it is singular, a similar analysis to that of Newcomb[9] for the single variable case can be applied.

$$\deg_{P_i} [Z_R] \leq \deg_{P_i} [Z] \qquad i=1,2,\ldots,n.$$

In particular if $Z(a, P_2,\ldots,P_n) = -Z(-a,P_2,\ldots,P_n)$,

$$\deg_{P_1} [Z_R] = \deg_{P_1} [Z] - 1$$

V. CONCLUSION

An extension to Richards' theorem for positive real matrices of several variables is shown, which will be very useful in the synthesis of noncommensurate transmission line n ports. It will be useful to arrive at some theorems which guarantee that $Z(1,P_2,\ldots,P_n)$ is a constant matrix K independent of $P_2,\ldots, P_n$, as was shown by Saito[2] for the multivariable positive real functions.

REFERENCES

1. H. Ozaki and T. Kasami, "Positive Real Functions of Several Variables and their Application to Variable Networks", IRE Trans. Circuit Theory, Vol. CT-7, pp. 251-260, September 1960.

2. M. Saito, "Synthesis of Transmission Line Networks by Multivariable Techniques", Symp. on Generalized Networks (Polytechnic Institute of Brooklyn), pp. 353,392, April 1966.

3. D. C. Youla, "The Synthesis of Networks Containing Lumped and Distributed Elements", Symp. on Generalized Networks (Polytechnic Institute of Brooklyn), pp. 289-343, April 1966.

4. Richards, "A Special Class of Functions with Positive Real Part in Half Plane", Duke Math. J., 14, 777-786 (1947).

5. Newcomb, R. W., "Nonreciprocal Transmission-Line n-Port Synthesis", Proceeding I.R.E.E., Australia, pp. 135-142, April 1965.

6. Bayard, M., Theorie des reseaux de Kirchhoff", Editions de la Revue d'Optique, Paris, 1954, 170.

7. Koga, T., "Synthesis of Finite Passive n-Ports with Prescribed Positive Real Matrices of Several Variables", Technical Report No. RADC-TR-67-365, Polytechnic Institute of Brooklyn, July 1967, also IEEE Trans. on Circuit Theory, pp. 1-23, March 1968.

8. Ayres, F., "Theory and Problems of Matrices", Schaum Publishing Co., New York, 1962, p. 134.

9. Newcomb, R. W., "Linear Multiport Synthesis", McGraw-Hill Book Co., New York, 1966, p. 137.

10. McMillan, B., "Introduction to Formal Realizability Theory II", Bell System Technical Journal, Vol. 31, May 1952, page 580.

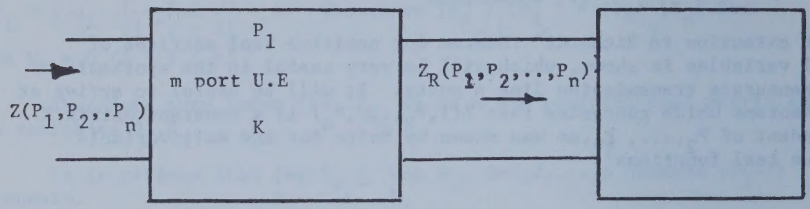


Figure 1. A block diagram representing an m port unit element extraction by Richards' Process.

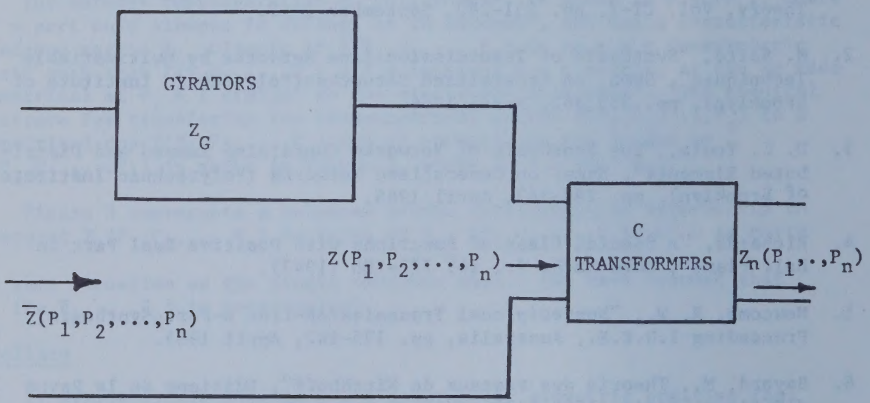


Figure 2. A block diagram (all blocks are m ports) representing the physical structure for transferring a nonsymmetrical matrix $\bar{Z}(P_1,P_2,\ldots,P_n)$ to a symmetrical one $Z(P_1,P_2,\ldots,P_n)$ and of normalizing the latter to $Z_n(P_1,P_2,\ldots,P_n)$.

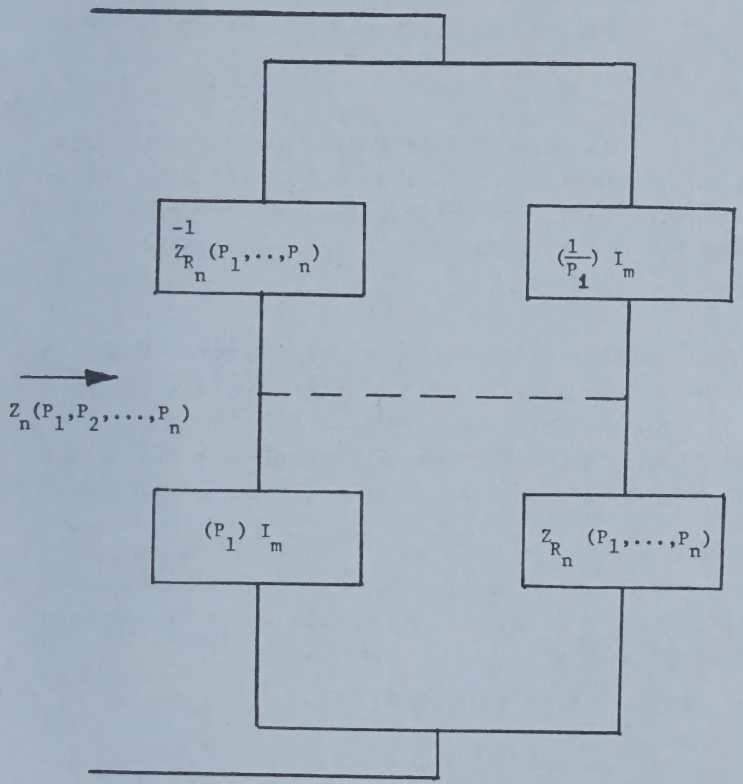


Figure 3. A block diagram (all blocks are m ports) representing a balanced bridge of Bott-Duffin, to interpret $Z_n(P_1,...,P_n)$ in terms of $Z_{R_n}(P_1,...,P_n)$.

REALIZATION OF A LOADED TWO-PORT WITH PRESCRIBED TRANSMISSION ZEROS AND DRIVING POINT FUNCTION

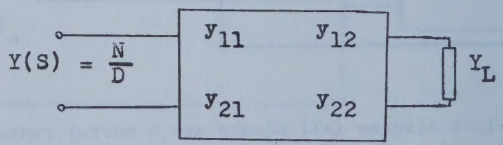
V.Ramachandran and M.N.S. Swamy
Department of Electrical Engineering
Sir George Williams Unuversity
Montreat, Quebec

Summary: In this paper, it is shown that any arbitrary driving point function can always be realized as a two-port active network, having prescribed transmission zeros in any one direction, and terminated by a prefixed passive RC admittance.

In the synthesis of arbitrary driving point functions, one of the methods adopted is to realize it as a two-port network terminated in an RC-impedance Z_L. In this method, the forward and backward zeros of transmission and Z_L are obtained only by the type of polynomial decomposition used.

In this paper, the problem of realizing an arbitrary driving point function with prescribed transmission zeros (either forward or backward) along with a prescribed passive RC load admittance will be examined.

For the structure shown below



Network M

we have $(y_{11}-Y)(Y_L-y_{22}) = y_{12}y_{21}$ ..(1)

Letting $Y_L = \frac{n_L}{d_L}$, $y_{11} = \frac{n_{11}}{d}$, $y_{12} = \frac{n_{12}}{d}$, $y_{21} = \frac{n_{21}}{d}$

and $y_{22} = \frac{n_{22}}{d}$

we can have

$$n_L d - d_L n_{22} = n_{21} D \qquad ..(2a)$$

and

$$n_{11} D - dN = d_L n_{12} \qquad ..(2b)$$

(2a) corresponds to TD (Th 3) with n_L and d_L as the two constituent parts of Horowitz decomposition. There is no unique solution, but this will enable one to determine the polynomials d and n_{22}.

From equation (2b), we can determine the polynomials n_{11} and n_{22} as shown below. Writing

$$\frac{n_{11}}{d_L} = K_{11}^{\infty} s + \Sigma \frac{K_{11}^{i} s}{s + d_i} + K_{11}^{o} \qquad ..(3)$$

where d_i's are the zeros of d_L, we get

$$K_{11}^{i} = \frac{-1}{d_i} \cdot \frac{n(-d_i)}{D(-d_i)} d(-d_i) \qquad ..(4)$$

From these residues, the polynomial n_{11} will be known. Hence, n_{12} is also uniquely determined. These completely specify the y-parameters of the network M, which can always be realized using a generalized gyrator and passive RC one-ports.

Thus the given specifications of $\frac{N}{D}$, $\frac{n_L}{d_L}$ and n_{21} are always satisfied by the network synthesized by the above method.

SYNTHESIS OF TRANSFER FUNCTIONS USING DIFFERENTIAL INPUT AMPLIFIERS

K.SANKARA RAO and V.G.K.MURTI
Department of Electrical Engineering
Indian Institute of Technology
Madras, India.

ABSTRACT

A new RC network configuration using grounded capacitors to realize any stable biquadratic voltage ratio transfer function is presented. The active element utilized is an integrated differential-input dual-output operational amplifier. The sensitivities of poles and zeros with reference to the passive elements are found to be quite low. To illustrate the procedure, a notch filter and a network with positive real zeros of transmission are synthesized.

1. INTRODUCTION

Even though the use of an amplifier for the realization of a transfer function by an active RC network has been known for a long time (Reference 1), the interest in this field has increased in recent years due to the availability of inexpensive and reliable linear operational amplifiers in integrated circuit form. A large number of methods (e.g., those in References 2 -5) are now available for the realization of a given voltage ratio transfer function - hereinafterwards, simply called the transfer function - by a grounded RC 2-port network in which is embedded one operational amplifier. While the conditions of applicability vary from one method to another, a common feature of all these methods is that no special provision exists to ensure that all the capacitors in the RC network are grounded - a characteristic that would facilitate adoption of the active configuration to integrated circuit construction (Reference 6).

In this paper, we present a configuration using an RC grounded network incorporating a differential-input dual-output operational amplifier to realize a general biquadratic transfer function of a stable network. The number of capacitors required is canonical (two) and both the capacitors are grounded.

It is well known that from the point of view of sensitivity of poles and zeros it is preferable to realize a higher order

transfer function by the cascade connection of subnetworks each realizing at most a biquadratic transfer function. Since in our configuration, the output terminals of the amplifier also serve as the output terminals of the 2-port subnetwork, the cascading of such stages without undesirable loading effects is readily accomplished.

2. STUDY OF A BIQUADRATIC FUNCTION

We will be concerned in this paper only with biquadratic transfer functions of a stable network. We first establish some properties of such a function which are of interest in our work.

After suitable frequency scaling, the transfer function can be brought to the form:

$$G(s) = H\frac{1+as+bs^2}{1+cs+s^2} = H\frac{N(s)}{D(s)} = HG_1(s) \; . \qquad (1)$$

While the constants H, a and b can be either positive or negative, stability considerations require that c be always positive.

2.1. Real Poles :

When the poles of G(s) are real and negative, the variation of the function along the negative real axis of the s-plane will be as in one of the typical curves shown in Figs. 1a, 1b and 1c. It can be observed from the curves that there always exists an interval on the negative real axis of the s-plane in which

either (a) $G_1(s) > 1$; $dG_1/ds > 0$

or (b) $0 < G_1(s) < 1$; $dG_1/ds < 0$.

2.2. Complex Poles :

We shall now prove that $G_1(s)$ cannot be monotonic in the interval $(-\infty,0)$ of s when G(s) has complex poles.

$$G_1(s) = \frac{1 + as + bs^2}{1 + cs + s^2} \; .$$

Effecting a change of variable, $s = x - c/2$, we get

$$G1(s) = f(x) = \frac{1 + a(x-c/2) + b(x-c/2)^2}{x^2 + (1-c^2/4)} \; .$$

$$dG_1(s)/ds = df(x)/dx = \frac{(bc-a)x^2 + (\overline{b-1}\,2 + \overline{a-b}\,c)x + (1-c^2/4)(a-bc)}{x^2 + (1-c^2/4)} \qquad (2)$$

A stationary point exists where df/dx is zero. As the poles of $G_1(s)$

are complex, $c<2$ and $(1-c^2/4)>0$. Therefore the coefficient of x^2 and the constant term in the numerator of the expression in (2) have opposite signs. Consequently, $df(x)/dx$ has one negative real zero, i.e., $G_1(s)$ has a maximum or minimum for a finite real value of s which is less than $(-c/2)$. Thus $G_1(s)$ is not monotonic in the range $(-\infty, 0)$ of s. The characteristic variation of $G_1(s)$ for different values of maximum and minimum is sketched in Figs. 1d -1f. From these, we note that there always exists an interval on the negative real axis of the s-plane where one of the conditions below is satisfied:

(a) $G_1(s)>1$; $dG_1/ds>0$,

(b) $0<G_1(s)<1$; $dG_1/ds<0$,

(c) $G_1(s)<0$; $dG_1/ds<0$.

3. BASIC NETWORK CONFIGURATION

The basic network configuration is as shown in Fig. 2, where the amplifier A is a differential-input dual-output amplifier satisfying the relation:

$$V_o = A(V_1 - V_2) = -V_3 \tag{3}$$

We assume that the 5-port network N has no internal nodes and that it contains only RC elements with all capacitors grounded. Hence the short-circuit transfer admittance function between any two ports is a constant, independent of s.

If now, the transfer admittance between ports 1 and 2 is zero, the current equations at nodes 1 and 2 become

$$0 = -g_{1i}V_i + y_{11}V_1 - g_{13}V_3 - g_{1o}V_o \tag{4a}$$

$$0 = -g_{2i}V_i + y_{22}V_2 - g_{23}V_3 - g_{2o}V_o \quad , \tag{4a}$$

where g_{pq} is the conductance of the element connecting the nodes p and q and y_{11} and y_{22} are the short-circuit driving-point admittances of N at ports 1 and 2.

Equations 3, 4a and 4b lead to the following expression for the transfer function of the overall configuration:

$$G(s) = \frac{V_o(s)}{V_i(s)} = \frac{y_{11}g_{2i} - y_{22}g_{1i}}{y_{11}(g_{23}-g_{2o}) - y_{22}(g_{13}-g_{1o}) - y_{11}y_{22}/A} \tag{5}$$

As $A \to \infty$, the expression (5) reduces to

$$G(s) = \frac{k_1 y_{11} - k_2 y_{22}}{k_3 y_{11} - k_4 y_{22}} , \qquad (6)$$

where k_1 and k_2 are positive while k_3 and k_4 may be either positive or negative.

4. REALIZATION

We have observed earlier in Section 2 that a general biquadratic function can after frquency scaling be put in the form,

$$\frac{N(s)}{D(s)} = H \frac{1 + as + bs^2}{1 + cs + s^2} ,$$

where H can take either positive or negative values. However, since our network configuration permits us to take the output from terminal 3 of the amplifier instead of from terminal o where required, we can assume without loss of generality that H is positive in the following discussion. If H indeed is negative, the negative of the required transfer function should be realized by the method to be described and the output taken from terminal 3 instead of terminal o.

Dividing the numerator and denominator by $(1+T_1 s)(1+T_2 s)$, a polynomial with distinct negative real zeros at $(-1/T_1)$ and $(-1/T_2)$, the transfer function can be expressed in the following form:

$$G(s) = H \frac{1 + a_1 s/(1+T_1 s) - a_2 s/(1+T_2 s)}{1 + b_1 s/(1+T_1 s) - b_2 s/(1+T_2 s)} . \qquad (7)$$

The values of T_1 and T_2 are chosen such that a_1, a_2, b_1 and b_2 are positive. It can be shown that this choice is always possible. For example, for positive values of b, the following choice of T_1 and T_2 leads to the positive values of the above constants:

$$1/T_1 > 1/T_2 > \begin{cases} a/b \\ c \end{cases} . \qquad (8)$$

The expression in (7) is now compared with (6). In the usual active network synthesis procedures, the terms with positive residues are associated with y_{11} and those with negative residues with y_{22}. However, the proposed configuration of N requires that both y_{11} and y_{22} should contain constant parts. Accordingly, on comparing the numerators of (7) and (6), we choose k_1, k_2, y_{11} and y_{22} as follows:

$$k_1 = k_2 = k_o H \qquad (9a)$$

$$y_{11} = 1 + g + a_1 s/(1+T_1 s) \qquad (9b)$$

$$y_{22} = g + a_2 s/(1+T_2 s) \qquad (9c)$$

where k_o and g are constants to be determined later.

Now, substituting the foregoing expressions for y_{11} and y_{22} in the denominator of (6) and equating the latter to

$$k_o [1 + b_1 s/(1+T_1 s) - b_2 s/(1+T_2 s)] ,$$

we have the following conditions to be fulfilled:

$$k_3 a_1 = k_o b_1 \qquad (10a)$$

$$k_4 a_2 = k_o b_2 \qquad (10b)$$

$$k_3(1+g) - k_4 g = k_o \qquad (10c)$$

Choosing k_3 and k_4 in accordance with (10a) and (10b) the following expression for g is obtained from (10c)

$$g = \frac{1 - (b_1/a_1)}{(b_1/a_1) - (b_2/a_2)} = \frac{a_2}{b_2} \cdot \frac{(a_1/b_1) - 1}{(a_2/b_2) - (a_1/b_1)} \qquad (11)$$

The choice of $(1/T_1)$ and $(1/T_2)$ is to be made to make g positive while at the same time ensuring that a_1, a_2, b_1 and b_2 are positive. If $N(s)$ and $D(s)$ are positive for $s = (-1/T_1)$ and $s = (-1/T_2)$, the conditions to be fulfilled are :

$$T_2 > T_1 \qquad (12)$$

and

$$\begin{cases} G_1(-1/T_2) > G_1(-1/T_1) > 1 & (13a) \\ \text{or} \\ 0 < G_1(-1/T_2) < G_1(-1/T_1) < 1 , & (13b) \end{cases}$$

since (a_1/b_1), (a_2/b_2) and 1 are equal to $G_1(-1/T_1)$, $G_1(-1/T_2)$ and $G_1(0)$ respectively.

The conditions (12) and (13a) can be fulfilled by choosing the values of $(-1/T_1)$ and $(-1/T_2)$ in the interval on the negative real axis of s-plane where $dG_1/ds > 0$ and $G_1(s) > 1$. Similarly the conditions (12) and (13b) could be satisfied by the appropriate choice of the values of $(-1/T_1)$ and $(-1/T_2)$ from the interval where $dG_1/ds < 0$ and $G_1(s) < 1$. From the discussion in Section 2 it is seen that either of the two sets of conditions can always be met except in the special situation where only case (c) of Section 2.2 is applicable. Even in the latter situation an appropriate choice of T_1 and T_2 can be made, as shown in the following paragraph.

The special situation referred to in the foregoing paragraph arises when one zero of $G_1(s)$ is on the negative real axis and a stationary point corresponds to a minimum value of $G_1(s)$, which is

negative, as shown by the curve in Fig. 1f. In this case, we choose the values of $(-1/T_1)$ and $(-1/T_2)$ in the region where N(s) is negative. D(s) of course is positive on the entire negative real axis of the s-plane as its zeros are complex conjugate. Taking $T_2 < T_1$, we now get

$$\frac{N(s)}{(1+T_1 s)(1+T_2 s)} = 1 + \frac{a_1 s}{1+T_1 s} - \frac{a_2 s}{1+T_2 s} \tag{14}$$

$$\frac{D(s)}{(1+T_1 s)(1+T_2 s)} = 1 - \frac{b_1 s}{1+T_1 s} + \frac{b_2 s}{1+T_2 s} \quad , \tag{15}$$

where a_1, a_2, b_1 and b_2 are positive.

Referring to the numerator of (6), the values of k_1, k_2, y_{11} and y_{22} are chosen as follows:

$$k_1 = k_2 = k_o H$$
$$y_{11} = 1 + g + a_1 s/(1+T_1 s)$$
$$y_{22} = g + a_2 s/(1+T_2 s)$$

Substituting these values in the denominator of (6) and comparing with (15) yields the relations:

$$k_3 a_1 = -k_o b_1 \tag{16a}$$
$$k_4 a_2 = -k_o b_2 \tag{16b}$$
$$k_3(1+g) - k_4 g = k_o \quad . \tag{16c}$$

Substituting the values of k_3 and k_4 from (16a) and (16b) into (16c) we get

$$g(b_2/a_2 - b_1/a_1) = 1 + b_1/a_1 \quad . \tag{17}$$

For g to be positive, the necessary condition is:

$$a_2/b_2 < a_1/b_1 \quad . \tag{18}$$

Since $a_2/b_2 = -G_1(-1/T_2)$ and $a_1/b_1 = -G_1(-1/T_1)$, (18) is equivalent to:

$$G_1(-1/T_1) < G_1(-1/T_2) \tag{19}$$

Hence the region of negative real axis of the s-plane in which $G_1(s)$ and dG_1/ds are both negative also satisfies the necessary condition for a positive g.

We thus note that every stable biquadratic function can be realized as a voltage ratio transfer function utilizing the network configuration presented here. After calculating g, we choose an appropriate value of k_o such that the constant parts of y_{11} and y_{22} are consistent with the values of the transfer conductances at the ports. Table I illustates the choice of the element values in the network.

TABLE I

Quantity	Value	
	(a) k_3, k_4 positive	(b) k_3, k_4 negative
k_o	smaller of the two values: $g/(H+b_1/a_1)$ and $(1+g)/(H+b_2/a_2)$	
$g_{2i} = g_{1i}$	k_oH	k_oH
g_{23}	k_ob_1/a_1	0
g_{2o}	0	$k_o\ b_1/a_1$
g_{1o}	0	$k_o\ b_2/a_2$
g_{13}	k_ob_2/a_2	0
y_{14}	$1 + g - k_o(H+b_2/a_2) + a_1s/(1+T_1s)$	
y_{24}	$g - k_o(H+b_1/a_1) + a_2s/(1+T_2s)$	

In the above table, y_{14} represents the admittance connecting node 1 to ground. Similarly y_{24}. The values of conductances and capacitors should be considered as normalized values so as to make $y_{11}(0) - y_{22}(0) = 1$. The scaling of all admittances by a suitable factor is of course always possible.

Example 1 :

We consider the following transfer function of a notch filter:

$$G(s) = (1+s^2)/(1+0.1s+s^2) = N(s)/D(s)$$

The maximum value of G(s) for negative real s occurs at s = -1. Accordingly, we have, on dividing N(s) and D(s) by (1+s)(1+0.1s) and expressing G(s) as in (7)

$a_1 = 101/90$; $a_2 = 20/9$; $b_1 = 100/90$; and $b_2 = 19/9$. Applying (11), g is calculated as 0.247. The element values of the network can now be obtained by using Table I.

$k_o = 0.124$; $g_{1i} = g_{2i} = 0.124$; $g_{23} = 0.123$; $g_{13}=0.118$; $g_{20}=g_{10}=0$. $y_{14} = 1.005 + (101/90)s/(1+0.1s)$ and $y_{24} = (20/9)s/(1+s)$.

Fig.3 depicts the network realizing the notch filter.

Example 2:

The transfer function to be realized is:

$$G(s) = (1-s)(1-2s)/(1+2s)(1+0.5s).$$

The values of $(-1/T_1)$ and $(-1/T_2)$ can be selected from the negative real axis to the left of the pole at -2 (see Fig. 1). We can choose $(1/T_1)$ as ∞ and $(1/T_2)$ as 2. The values of constants in the partial fraction expansion according to (7) are:

$a_1 = 4$; $a_2 = 7.5$; $b_1 = 2$; and $b_2 = 0$.

Accordingly, the element values are:

$k_0 = 2/3$; $g_{1i} = g_{2i} = 2/3$; $g_{23} = 1/3$; $g_{13} = g_{20} = g_{10} = 0$;

$y_{14} = (4/3) + 4s$; and $y_{24} = 7.5s/(1+0.1s)$.

The circuit is shown in Fig. 4.

5 SENSITIVITY CONSIDERATIONS

It can be shown that the sensitivities of the poles and zeros of the transfer function with respect to the variation of A is inversely proportional to A as in many other methods using operational amplifiers and can be considered to be negligibly small.

The sensitivity of the magnitudes of zeros with respect to variation of passive components can be calculated as follows for condition (a) of Table I. The zeros of G(s) are the roots of the numerator of (5), i.e., the roots of the equation:

$$g_{2i}\left[g_{1i} + g_{13} + g_{14} + (a_1 s/(1+T_1 s))\right] - g_{1i}\left[g_{2i} + g_{23} + g_{24} + a_2 s/(1+T_2 s)\right] = es^2 + fs + 1 = 0,$$

where g_{14} and g_{24} are the conductances connecting directly nodes 1 and 2 respectively to ground. The values of e and h are:

$$e = T_1 T_2 \left[g_{2i}(g_{13} + g_{14} + a_1/T_1) - g_{1i}(g_{23} + g_{24} + a_2/T_2)\right] \text{ and}$$

$$h = g_{2i}(g_{13} + g_{14}) - g_{1i}(g_{23} + g_{24}).$$

Observing that the nominal values of e and h are bg_{2i} and g_{2i}, the following sensitivity figures can be calculated where xSy stands for the the sensitivity of x with respect to y.

$eSg_{2i} = (g_{13} + g_{14} + a_1/T_1)T_1 T_2/b$; $eSg_{1i} = (g_{23} + g_{24} + a_2/T_2)T_1 T_2/b$;

$hSg_{2i} = g_{13} + g_{14}$; $hSg_{1i} = g_{23} + g_{24}$; $eSg_{13} = T_1 T_2 g_{13}/b$; $hSg_{13} = g_{13}$

$eSg_{23} = T_1 T_2 g_{13}/b$; $hSg_{23} = g_{23}$; $eSg_{14} = T_1 T_2 g_{14}/b$;

$hSg_{14} = g_{14}$; $eSg_{24} = T_1 T_2 g_{24}/b$; and $hSg_{24} = g_{24}$.

Since the conductances g_{13}, g_{23} etc. are normalized values the above expressions for sensitivity can be seen to be dimensionless.

If $w_o^2 = h/e$, we get the sensitivity figures for w_o^2 as

$$w_o^2 Sg_{2i} = |hSg_{2i} - eSg_{2i}| \text{ etc.}$$

The values of the sensitivities of w_o^2 with respect to various passive components for Example 1 considered in Section 4 are given below.

$w_o^2 Sg_{1i} = w_o^2 Sg_{2i} = 0.1105$; $w_o^2 Sg_{13} = 0.1062$; $w_o^2 Sg_{23} = 0.1107$;

$w_o^2 Sg_{14} = 0.9045$; $w_o^2 Sa_1 = 1.112$; and $w_o^2 Sa_2 = 0.222$

Similarly the sensitivity for poles of G(s) and that for conditions of case (b) of Table I can be calculated. In the above example the sensitivities of poles are of the same order as those for zeros as k_3 and k_4 are nearly equal to k_o.

The numerator of G(s) involves the difference of two terms, $k_o H(1+g)$ and $k_o Hg$. Similarly the denominator involves the difference of $k_3(1+g)$ and $k_4 g$. To minimize the variation of the resulting term, due to variation in element conductances, it is desirable to have as low a value for g as possible, keeping the element spread in mind. To achieve a low value of g, it can be seen on reference to (11) that T_1 and T_2 should be so chosen that $G_1(-1/T_1)$ is close to unity and far removed from $G_1(-1/T_2)$.

6. CONCLUSIONS

Even though a number of RC active network synthesis schemes have been published, the following specific advantages of the procedure suggested in this paper merit attention as they do not appear to be simultaneously present in any single-amplifier RC network realization method suggested so far.

1. It is applicable to any general biquadratic function with left half-plane poles.

2. The number of capacitors required is minimal, thus providing advantages from the point of view of tuning to different frequencies and avoiding difficulties in alignment. Furthermore, both the capacitors are grounded which is a useful feature in integrated construction.

3. Cascading of different subnetworks to realize a higher order transfer function is readily permitted.

The sensitivity figures are comparable with those for previously given single amplifier configurations.

REFERENCES

1. Sallen,R.P. and Key,E.L., IRE Tran. Circ. Theory, CT-2, March 1955.
2. Pande,H.C. and Shukla,R.S., Proc. IEE, vol.112, Dec. 1965.
3. Holt,A.G.J. and Sewell,J.I., Proc. IEE, vol. 112, Dec. 1965.
4. Venkataramani,K. and Murti,V.G.K., Intl. Jour. of Electronics, vol. 23, Oct. 1967.
5. Pande,H.C. and Shukla,R.S., Jour. Inst. Tele-comm. Engrs. (New - Delhi), vol. 13, Feb. 1967.
6. Kerwin,W.J.,Huelsman,L.P. and Newcomb,R.W., IEEE Jour. Solid-State Circuits, vol. SC-2, Sep. 1967.

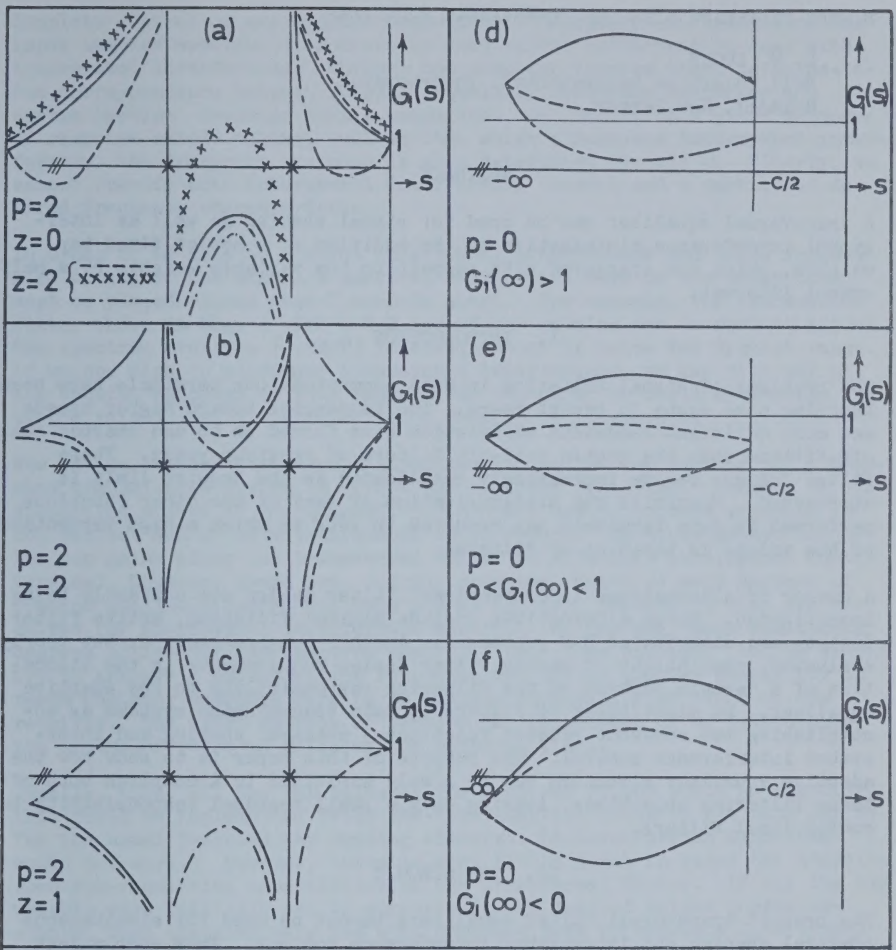


FIG.1 p= no. of negative real poles; z = no. of negative real zeros.

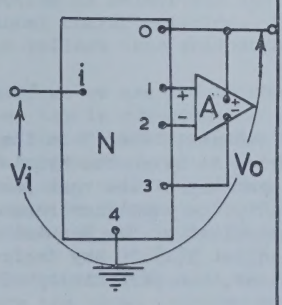


FIG.2

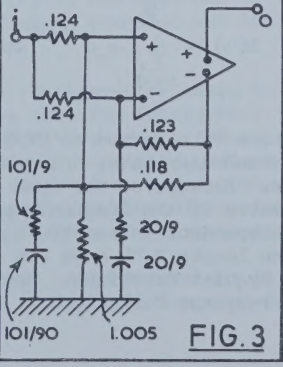


FIG.3

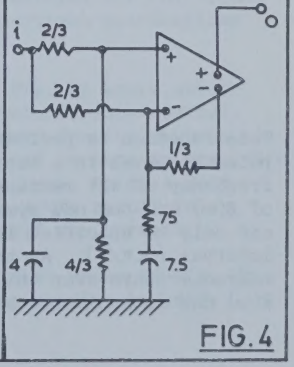


FIG.4

SIGNAL FILTERING WITH THE TRANSVERSAL EQUALIZER

R. W. LUCKY
Bell Telephone Laboratories, Incorporated
Holmdel, New Jersey

ABSTRACT

A transversal equalizer can be used for signal shaping as well as intersymbol interference minimization by the addition of a set of fixed tap weights, which are staggered with respect to the variable weights by a half-symbol interval.

I. INTRODUCTION

The problems of signal filtering in data communications terminals have been becoming more acute in recent years. The tendencies toward higher speeds and more efficient bandwidth utilization have forced us to use sharper roll-off filters than the gentle roll-off filters of previous years. These filter designs become increasingly complicated as the Nyquist limit is approached. Meanwhile the miniaturization of many of the other functions performed in data terminals has resulted in sets in which a high percentage of the volume is taken up by filters.

A number of alternatives to conventional filter design are currently being investigated. These alternatives include digital filtering, active filter design, and distributed R-C synthesis. Another long-recognized, but poorly evaluated, possibility of easing filter design requirements is the allocation of a certain portion of the filtering responsibility to the adaptive equalizer. We might think of filters in data transmission systems as accomplishing two somewhat related functions - spectral shaping and intersymbol interference control. The purpose of this paper is to show how the adaptive equalizer structure can be simply exploited to accomplish both of these filtering objectives, leaving only a small residual responsibility to conventional filters.

II. THE CIRCUIT

The present transversal filter equalizers cannot be used for simultaneous spectral shaping and intersymbol interference control. They simply lack the necessary degrees of freedom to accomplish both objectives. For a delay line with tap weights c_j, $j = -N,\ldots,+N$, the transfer function is

$$E(\omega) = \sum_{n=-N}^{N} c_n e^{-in\omega T}. \qquad (1)$$

This function is periodic with a period of $2\pi/T$ radians. When T is the pulse interval in a baseband data wave, this period is twice the Nyquist frequency of π/T radians. Because of the even symmetry of the real part of $E(\omega)$ and the odd symmetry of its imaginary part, the equalizer response can only be specified independently over the equivalent of one Nyquist interval $[0,\pi/T]$. As the length $N \rightarrow \infty$ we can adjust $E(\omega)$ to any desired characteristic over the Nyquist bandwidth. However, the periodicity of $E(\omega)$ now determines its response for all ω.

Complete control of one Nyquist bandwidth is necessary and (providing the input sampled spectrum is nonzero in this range) sufficient to deal with intersymbol interference. Through the sampling theorem there is a one-to-one correspondence between possible equalizer frequency responses and system impulse response sample sequences. By specifying a given sequence of response sample values, such as that which eliminates intersymbol interference, the frequency response is also determined for all ω. Clearly, we cannot specify both intersymbol interference removal and a particular desired frequency characteristic.

In order to be able to control intersymbol interference and leave freedom also to adjust the signal's spectral shaping the taps on the delay line must be placed closer than T seconds apart. For example, the transversal filter shown in Fig. 1 with a T/2 second tap spacing can be used to adjust the spectrum over the $[0,2\pi/T]$ interval, which is twice the Nyquist range. If we now wish to eliminate intersymbol interference, we can pick any of the Nyquist shapings in this range. For instance, we could synthesize any of the raised cosine characteristics from 0- to 100-percent roll-off.

Now let us suppose that such a structure is used at the receiver of a baseband transmission system to effect the receiver filtering. Because of the distorting effect of the channel, the overall system impulse response is not perfect and we have intersymbol interference. We could adjust all of the tap gains along the transversal filter to alleviate intersymbol interference. However, there are, roughly speaking, twice as many degrees of freedom available (since the taps are at T/2 sec. intervals) than are required for intersymbol interference control. We can use these additional degrees of freedom for spectrum control. This does not mean that we should try to achieve a raised cosine response, but rather that we should try to reach whatever response results in minimum probability of error at the receiver. The extra degrees of freedom can be thought of as determining the amount of received noise at the sampling times.

Ideally, then, all the tap gains at the receiver filter would be adaptive. This could be implemented using RMS-type adaptive loops in the same way as for the usual T-second tap spacing filters. (A zero-forcing algorithm would not work.) However, there is very little point in using the adaptive spectrum-regulating capabilities of the transversal filter. If all the tap weights were initially set to approximate the desired raised cosine response, then changing only the even numbered taps to regulate intersymbol interference would be very close to optimal. In the telephone environment of high signal-to-noise ratio the control of intersymbol interference is crucial and must be precise. On the other hand the control of the noise spectrum is relatively uncritical and need not be tailored for individual channel characteristics. Present practice and theoretical calculations both reflect this philosophy.

Therefore we are led to the configuration shown in Fig. 2a where every other tap is adaptive and is used for intersymbol interference control. Here the adaptivity could be either of the RMS or zero-forcing variety. The intermediate taps are simply constants, determined perhaps by precision resistors, which have been computed on a compromise basis to give good average performance. Most probably the calculation would result in a nominal raised cosine spectrum when the channel itself was perfect.

The circuit of Fig. 2a is completely equivalent to that shown in Fig. 2b, where the usual transversal filter with T-second tap spacing is used.

Now we have simply taken a conventional adaptive equalizer and added a resistance matrix, possibly another summing amplifier (or more) and a T/2 second delay network. The resulting configuration has the same intersymbol interference control capabilities as before, but it now has the additional power to do much of the spectral control generally associated with conventional receiver filters.

The principle does not need to be restricted to the configuration of Fig. 3b where a band of width $2\pi/T$ is regulated. By taking two resistance branches from each tap on a conventional equalizer, passing one set of resistance branches through a T/3 delay and the other through a 2T/3 delay, a bandwidth of $3\pi/T$ can be controlled. More taps can be added for further control, but the basic circuit of Fig. 3b would probably be suitable for most data applications.

An equivalent block diagram of the circuit is shown in Fig. 3. The upper branch contains an ideal filter of the raised cosine or whatever other form might be desired, while the lower branch contains a transversal filter of the conventional T-second tap spacing variety. The tap gains of this lower branch are designated ε_j, $j = -N,\ldots,+N$, and where ε_j is the perturbation added or subtracted from the ideal jth tap gain in order to minimize intersymbol interference on the particular channel in use. When the channel is perfect, the gains ε_j are zero and only the upper branch is in use.

In order to evaluate this system we need to compare it with a conventional filter-equalizer series combination with respect to noise performance and intersymbol-interference handling capability. In the following sections we first examine its noise performance and then move on to an evaluation of a specific application - that of simplifying the conventional VSB receiver by removal of all bandpass filtering.

III. IDEAL PERFORMANCE OF THE CIRCUIT

We shall now suppose that the transversal filter is infinitely long. For a perfect channel the response is exactly that of a raised-cosine pulse and differs in no respect from that of the conventional filter-equalizer tandem combination.

We now introduce a channel containing parabolic delay distortion. The parabolic delay has magnitude β_m in normalized pulse intervals at the Nyquist frequency. An ideal receiver would not be degraded in noise performance by this channel at all. (The distortion is completely removed with no enhancement of noise by an all-pass filter.) However, the conventionally equalized system is degraded in noise performance by a small amount. Since we have assumed an infinite equalizer the intersymbol-interference can be completely removed. However, in so doing, the equalizer slightly augments the received noise power. The amount has been previously calculated in Reference 1.

Block diagrams of the conventional and proposed systems (baseband equivalents) are shown in Fig. 4. In each case the filter $G(\omega)$ represents a square-root-of-Nyquist frequency response so that $|G(\omega)|^2$ is a perfect Nyquist characteristic. The basic difference between the two systems is in the placement of the transversal filter, i.e., whether it is used in series or in parallel with the receiving filter $G(\omega)$.

In the conventional system the equalizer response which completely eliminates intersymbol interference for sampling at time zero is given by

$$E_1(\omega) = \frac{1}{[\,|\,G(\omega)\,|^2 C(\omega)\,]_{eq}} \tag{2}$$

where $C(\omega)$ is the channel response and the notation

$$[X(\omega)]_{eq} = \sum_k X\left(\omega + \frac{2k\pi}{T}\right) \tag{3}$$

indicates the "equivalent" Nyquist bandwidth response. (Equivalent in the sense that its samples at times nT are identical.) Since the sampling is at time zero it is necessary to modify $C(\omega)$ by a time delay factor $e^{-i\omega t_0}$ to account for whatever actual reference time t_0 is used.

With the equalizer response, Eq. 2, the normalized noise variance at the received sampling times (for white channel noise) is

$$\frac{N_1}{N_0} = \frac{1}{\pi}\int_0^\infty \frac{|\,G(\omega)\,|^2}{\left|\,[\,|\,G(\omega)\,|^2 C(\omega)\,]_{eq}\,\right|^2}\,d\omega. \tag{4}$$

In other words N_1/N_0 represents the increase in noise power at the received sampling instants because of the introduction and equalization of the channel $C(\omega)$.

In the proposed parallel system the equalizer response for perfect intersymbol interference removal is given by

$$E_2(\omega) = \frac{1 - [\,|\,G(\omega)\,|^2 C(\omega)\,]_{eq}}{[G(\omega)C(\omega)]_{eq}}\,. \tag{5}$$

(This equation merely accounts for the fact that the Nyquist equivalent of the overall response must be constant.) This response leads to a noise enhancement factor of

$$\frac{N_2}{N_0} = 1 + \frac{1}{\pi}\int_0^\infty \left|\frac{1 - [\,|\,G(\omega)\,|^2 C(\omega)\,]_{eq}}{[G(\omega)C(\omega)]_{eq}}\right|^2 d\omega. \tag{6}$$

Because of the periodic nature of the integrand in Eq. 6 the ratio N_2/N_0 is infinite. However, we should expect that the noise is actually band-limited to not more than twice the Nyquist frequency. If this is not the case, then we should have a low-pass filter at the input of the receiver. In any event, we assume a pessimistic case of bandlimiting to $2\pi/T$ - this becomes the upper limit of integration in Eq. 6.

For our example we have chosen a channel given by

$$C(\omega) = e^{i(\beta_m T^3\omega^3/3\pi^2 - \omega t_0)} \tag{7}$$

(parabolic delay of β_m pulse intervals at the Nyquist frequency and reference delay t_0). In addition we use 50 percent raised cosine filtering, i.e.,

$$\begin{aligned} |G(\omega)|^2 &= T && \text{for} \quad 0 \le \omega \le \frac{\pi}{2T} \\ |G(\omega)|^2 &= \frac{T}{2}(1 + \sin \omega T) && \text{for} \quad \frac{\pi}{2T} < \omega \le \frac{3\pi}{2T} \\ |G(\omega)|^2 &= 0 && \frac{3\pi}{2T} < \omega \end{aligned} \tag{8}$$

Expressions 7 and 8 were substituted into Eqs. 4 and 6 and, after some algebraic simplification, evaluated by computer. For optimum reference times t_0 the following numerical results were obtained.

β_m	N_1/N_0 (db)	N_2/N_0 (db)	Comments
0	0	0	
1	.016	.119	usual distortion
2	.050	.88	moderately bad
3	.123	2.22	severe distortion
4	.191	2.25	
5	.255	2.41	

It can be seen that the parallel transversal filter system is about 1 db inferior to the conventional system for ordinarily encountered values of delay distortion.

IV. EVALUATION OF AN APPLICATION - A SIMPLIFIED VSB RECEIVER

For a specific application we evaluate the performance of a vestigial sideband (VSB) receiver using only Butterworth low-pass filters in addition to the modified equalizer. A block diagram of the receiver structure is shown in Fig. 5. In this diagram a signal consisting of amplitude

modulated pulses together with bandlimited white noise is the input to the receiver. The response of the system up to the input of the receiver is denoted $S(\omega)$.

The input signal at the receiver is passed through filter $H(\omega)$, which is normally a bandpass filter designed to accomplish the square root of the desired raised cosine pulse shaping. The signal is demodulated by a coherent carrier with zero phase and then filtered by low-pass filter $G(\omega)$. This low-pass filter removes the undesired modulation products as well as accomplishing some baseband shaping. Generally, since square pulses rather than impulses are used at the transmitter, a baseband shaping of the form $\omega/\sin\omega$ is required. The receiver filter $G(\omega)$ does "half" this baseband shaping, i.e., has a factor of the form $\sqrt{\omega/\sin\omega}$. Finally, the signal is passed through an N-tap transversal equalizer $T(\omega)$ and sampled at times nT.

In order to simplify this receiver structure, we shall substitute simple low-pass filters, intended only for noise suppression, for the bandpass filter $H(\omega)$ and for the low-pass-shaping filter $G(\omega)$. All the signal shaping requirements will be undertaken by the placement of intermediate taps on the transversal filter.

In our example we assume an ideal transmitter and channel, so that we are only concerned with the ability of the finite equalizer to assume the signal shaping responsibilities under quiescent conditions. When a real channel is substituted there will be some signal-to-noise degradation resulting from the equalizer's readjustment for intersymbol interference minimization. We already have some idea of this penalty from the results of the preceding section. In addition, the equalizer may not be able to suppress intersymbol interference to the extent possible in the unmodified receiver (or it may be able to do it better). This ability has not been investigated, but from a rudimentary study of the pulse response in the modified system, it would appear that the intersymbol interference suppression capability is relatively unimpaired.

The signal spectrum, $|S(\omega)|^2$, in the example was chosen to be that of a 50 percent raised cosine baseband system, modulated by a carrier and then passed through a 50 percent raised cosine vestigial sideband filter. This is somewhat typical of a practical VSB system. The received spectrum is symmetrical with the (removed) carrier position located 3/4 of the way across its base, from left to right.

Both low-pass filters, $H(\omega)$ and $G(\omega)$, were arbitrarily specified as 4th order Butterworth filters. The cutoff frequency for the initial filter $H(\omega)$ was taken at the carrier frequency, while the cutoff frequency for the demodulating filter $G(\omega)$ was placed at the edge of the demodulated baseband spectrum. With all the system now fixed except the transversal filter $T(\omega)$, we can optimize the quiescent tap values at T/2 intervals to determine minimum interference (noise and intersymbol interference) for a given length of equalizer. The mathematics of all this is perfectly straightforward, although somewhat tedious. An expression for mean square error in the output signal at sampling times is derived using the methods of References 1 or 2. This expression is differentiated with respect to the tap gains and set to zero, resulting in a set of simultaneous linear equations for optimum tap weights. (All the details have been omitted here in favor of brevity.) The resulting minimum RMS error is listed below as a function of the number of taps (spaced at T/2 intervals) for the noiseless case and for a 20 db signal-to-noise ratio.

Equalizer Taps (T/2 spacing)	RMS Error No Noise	RMS Error 20 db S/N
3	.1525	.3281
5	.0720	.1489
7	.0134	.1204
9	.0024	.1140
11	.0016	.1129
13	.0011	.1126
15	.0010	.1125
17	.0009	.1125

A short equalizer of 9 taps, which is only a 5-tap conventional equalizer augmented by 4 fixed intermediate taps, is seen to reduce the intersymbol interference to near negligible values. It is not clear whether such good results reflect primarily the utility of the proposed equalizer or of the all low-pass VSB receiver. However, a conventional 5-tap equalizer without the intermediate taps has an error about 15 times as great as the proposed system and much poorer noise rejection.

V. CONCLUSION

There is much to be gained by elimination of complicated filter requirements in data transmission terminals. We have suggested one way in which these requirements can be considerably eased. The proposed technique uses fixed gains at each tap of the automatic equalizer in addition to the already present variable gain. By means of one T/2 second delay all of the fixed gains are staggered with respect to the variable gains constituting the normal automatic equalizer. The resulting configuration can be used for spectral control of the data signal in addition to the usual equalization function.

On the plus side for the technique is its simplicity. The filter characteristics could be set by a precision resistance matrix and easily changed at any time. Also one of the times at which the filtering requirements become most severe is when a sharp roll-off system is designed. But this is exactly the time when a long automatic equalizer is required and many taps are available for this technique of spectral control. We have shown in an example design of a VSB receiver that only a few fixed gains might be required in the case of a 50 percent roll-off system. When sharper roll-off was required more taps would be needed and more generally would be available.

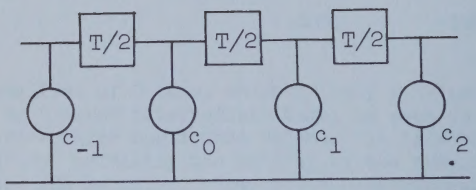
On the negative side, the proposed system lets in more noise than conventional methods. We evaluated this penalty for parabolic delay compensation and arrived at something less than a decibel of degradation. Also the proposed system may have less ability to compensate for intersymbol interference than the conventional usage of automatic equalization. This we did not attempt to evaluate, although in our example VSB system the correction ability of the automatic equalizer did not appear significantly impaired.

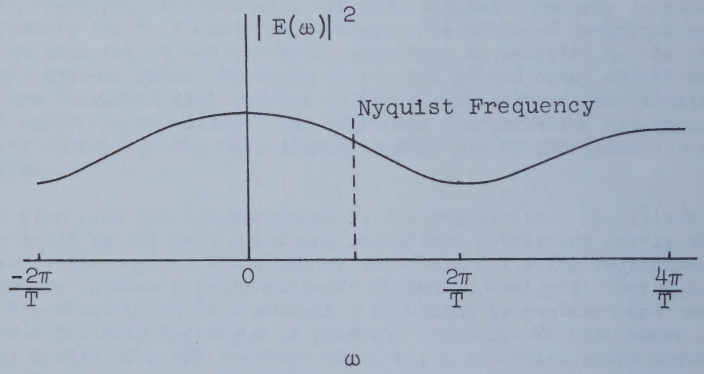
Finally, the convergence properties of a zero-forcing equalizer may be destroyed by the proposed configuration. Though the technique will work in general for rms-type control, the uncorrected pulse response may not satisfy convergence criteria for zero-forcing (our example is a case in point).

REFERENCES

1. Lucky, R. W., J. Salz and E. J. Weldon, Jr., Principles of Data Communication, Chapter 5, McGraw-Hill, New York, 1968

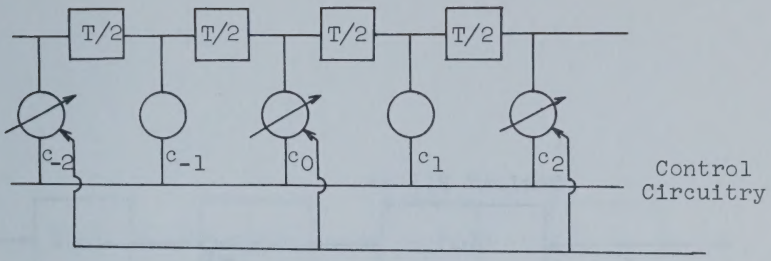
2. Gibby, R. A. and J. W. Smith, "Some Extension of Nyquist's Telegraph Theory," B.S.T.J., September, 1965.



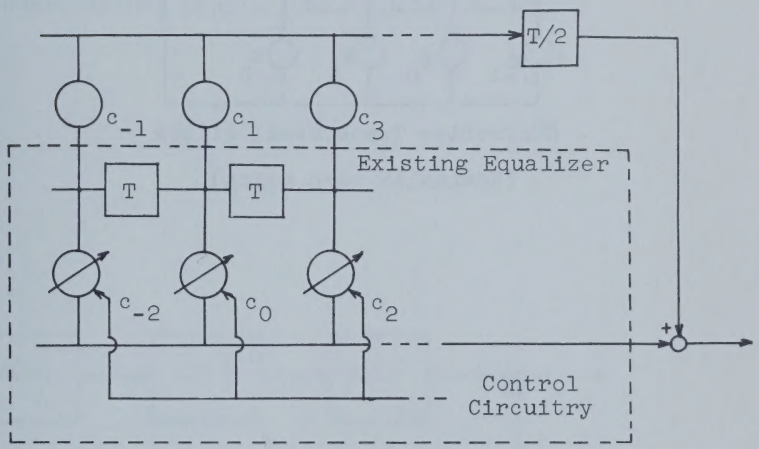


T/2 - Spacing Equalizer and its Frequency Response

FIGURE 1



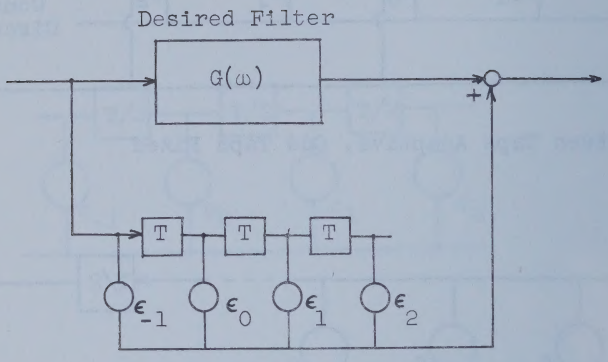
a. Even Taps Adaptive, Odd Taps Fixed



b. Modification of Existing Equalizer

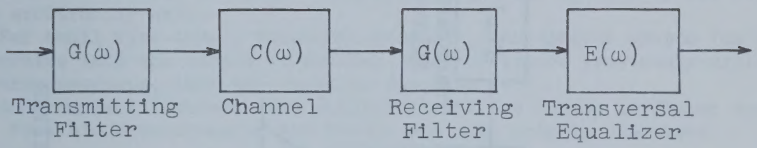
EQUALIZER CONFIGURATIONS

FIGURE 2

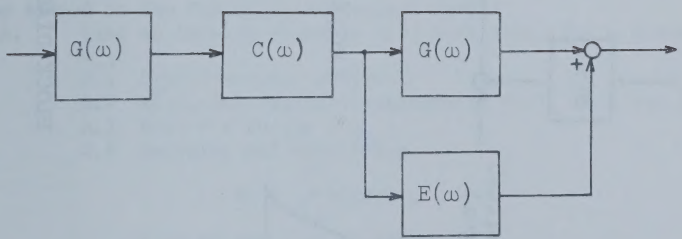


EQUIVALENT CIRCUIT

FIGURE 3



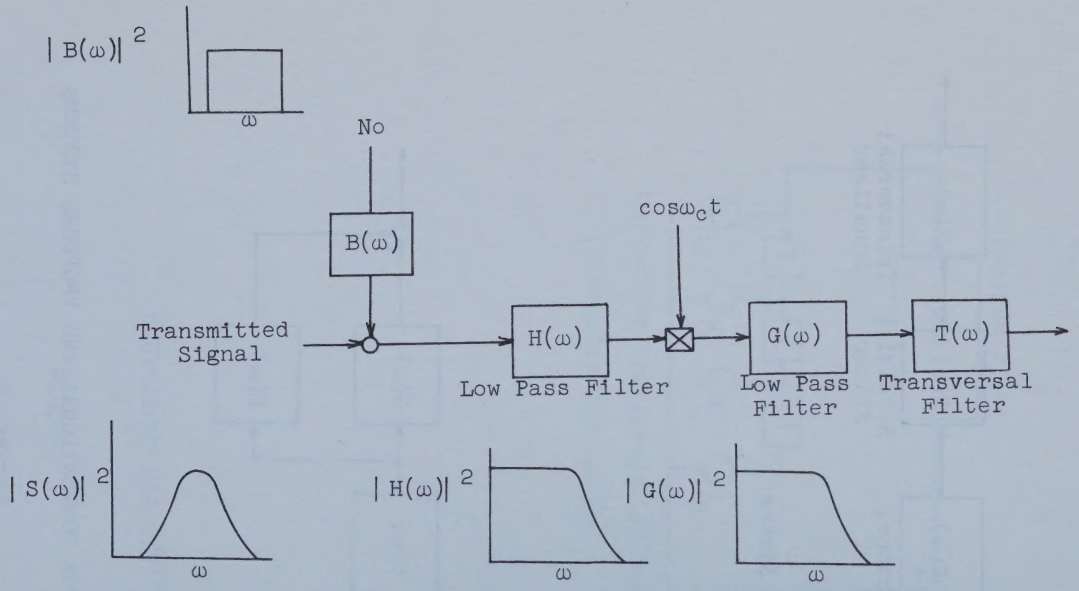
a. Conventional System



b. Proposed System

BLOCK DIAGRAMS FOR CONVENTIONAL AND PROPOSED SYSTEMS

FIGURE 4



BLOCK DIAGRAM OF ALL-LOW-PASS VSB RECEIVER

FIGURE 5

BANDLIMITED SYSTEMS WITH OPTIMUM TIME-RESPONSE

GABOR C. TEMES
University of California, Los Angeles and Ampex Corporation, Redwood City, California

ABSTRACT

System functions having the desirable properties of limited bandwidth, monotonic step response and the fastest possible rise-time are discussed. They turn out to be closely related to the square of one of the angular prolate spheroidal functions. From the system-function, the fastest x% to 100-x% rise-time can easily be found for any prescribed bandwidth and $\underline{x}$. Moreover, using existing circuit synthesis techniques, realizable transfer functions can readily be found which approximate the optimum system functions arbitrarily well.

For small rise-time x bandwidth products, the optimum system function degenerates into the simple triangular characteristics previously explored;[1] for large products, into the Gaussian function.

An example illustrates the application of the theory in filter synthesis. Possible extensions of the theory are also briefly discussed.

I. INTRODUCTION

A practical problem which frequently arises in the design of multilink video and data transmission systems is to find a reasonable compromise between the time- and the frequency-domain responses. A desirable time-response has no or negligible overshoot and the fastest possible rise-time; on the other hand, the corresponding frequency-response must be bandlimited to permit the use of practical components and also to suppress unwanted high-frequency signals.

Due to uncertainty limitations arising from the very definitions of time and frequency, the fastest rise-time under the described restrictions is a well-defined nonzero quantity. The purpose of this paper is to define the "optimum" system function which yields this fastest response. The asymptotic properties and applications of the theoretical results are also of interest and will be briefly discussed. Finally, some possible extensions of the presented theory are described.

II. SYSTEM FUNCTION

On the basis of the above considerations, the design problem at hand can be stated in the following equivalent forms:

A. Find an impulse-response h(t) such that, for a given bandwidth ω_o,

A.1 $0 \leq h(t) < \infty$ for $-\infty < t < \infty$;

A.2 $H(\omega)$, the Fourier-transform of h(t), is 1 for $\omega = 0$;

A.3 $H(\omega) = 0$ for $|\omega| > \omega_o$;

A.4 defining and specifying

$$\Delta a = \int_{-T}^{T} h(t)dt \ ,$$

$\underline{T}$ is as small as possible.

These requirements are illustrated in Figure 1.

B. Find a step-response a(t) such that, for a given bandwidth ω_o,

B.1 $0 \leq da(t)/dt < \infty$ for $-\infty < t < \infty$;

B.2 $\lim_{t_o \to \infty} [a(t_o) - a(-t_o)] = 1$;

B.3 $A(\omega)$, the Fourier-transform of a(t), satisfies

$$A(\omega) = 0 \quad \text{for } |\omega| > \omega_o \ ;$$

B.4 defining and specifying (Figure 2)

$$\Delta a = a(T) - a(-T),$$

$\underline{T}$ is as small as possible.

It is simple to show that requirement 4 in either formulation is equivalent to asking for maximum Δa for given $\underline{T}$. This latter requirement is somewhat easier to handle and will be considered hereafter.

Crucial to the solution of the problem is the following theorem of Krein:[2]

If $h(t)$ is an integral function of the exponential type, satisfying

$$0 \leq h(t) < \infty \quad \text{for} \quad -\infty < t < \infty$$

and bandlimited to $|\omega| \leq \omega_o$, then there exists a function $g(t)$ bandlimited to $|\omega| \leq \omega_o/2$ such that

$$h(t) = |g(t)|^2 \quad .$$

Accordingly, we may rephrase the original design problem in terms of $g(t)$:

C. Find a real or complex function $g(t)$ such that, for given ω_o,

C.1 $0 \leq |g(t)| < \infty$ for $-\infty < t < \infty$;

C.2 $\int_{-\infty}^{\infty} |g(t)|^2 \, dt = 1$;

C.3 $g(t)$ is bandlimited to $|\omega| \leq \omega_o/2$;

C.4 for a given $\underline{T}$, the quantity

$$\Delta a = \int_{-T}^{T} |g(t)|^2 \, dt$$

is as large as possible.

It is well known[3,4] that the solution of this problem is given by the renormalized angular prolate spheriodal functions $\psi_i(c,t)$. These are real functions of a parameter $c = \omega_o T/2$ as well as $\underline{t}$ and have the following relevant properties:

1. all $\psi_i(c,t)$ are bandlimited to $|\omega| \leq \omega_o/2$;

2. $\int_{-\infty}^{\infty} \psi_i \psi_j \, dt = \delta_{ij}$;

3. $\int_{-T}^{T} \psi_i \psi_j \, dt = \lambda_i(c) \delta_{ij}$, where $\lambda_o > \lambda_1 > \lambda_2 \ldots$

4. $\lambda_i \psi_i(c,t) = \int_{-T}^{T} \frac{\sin c(t-s)}{\pi(t-s)} \, \psi_i(c,s) ds$

5. $\mu_i c_i(c,t) = \int_{-T}^{T} e^{jcts} \psi_i(c,s) ds$.

Most importantly, the functions $\psi_o(c,t)$ and their eigenvalues $\lambda_o(c)$ satisfy the following theorem:[4]

For any function $g(t)$ bandlimited to $|\omega| \leq \omega_o/2$ and normalized to satisfy $\int_{-\infty}^{\infty} |g(t)|^2 dt = 1$, the inequality

$$\int_{-T}^{T} |g(t)|^2 dt \leq \lambda_o$$

holds. Equality is obtained only for $|g(t)| \equiv \psi_o(c,t)$ with $c = \omega_o T/2$.

Applying these results to the present design problem, it is evident that the desired optimum impulse-response is

$$h(t) = [\psi_o(t, \omega_o T/2)]^2 \quad .$$

The Fourier-transform of $h(t)$, $H(\omega)$, can also be found by noting that the Fourier-transform of $\psi_o(t)$ is a bandlimited and rescaled replica of $\psi_o(t)$:

$$\Psi_o(\omega_o,T,\omega) = \sqrt{\frac{4\pi T}{\omega_o \lambda_o}}\ \psi_o\left(c, \frac{2T\omega}{\omega_o}\right) \text{ for } |\omega| \leq \omega_o/2$$

$$\Psi_o(\omega_o,T,\omega) = 0 \quad \text{for} \quad |\omega| > \omega_o/2$$

Hence,

$$H(\omega_o,T,\omega) = \Psi_o(\omega_o,T,\omega) * \Psi_o(\omega_o,T,\omega) =$$

$$= \frac{2T}{\omega_o \lambda_o} \int_{1/2(-\omega_o+|\omega|+\omega)}^{1/2(\omega_o-|\omega|+\omega)} \psi_o\left(\frac{2Tu}{\omega_o}\right)\psi_o\left(\frac{2T(u-\omega)}{\omega_o}\right)du \text{ for } |\omega| \leq \omega_o$$

$$H(\omega_o,T,\omega) = 0 \quad \text{for} \quad |\omega| > \omega_o$$

The equations giving $h(t)$ and $H(\omega)$ constitute the analytical solution of the problem.

III. ASYMPTOTIC BEHAVIOR

The cases $T \to 0$ and $T \to \infty$ are of special interest. For the former, the fourth requirement of all three problem formulations (A,B and C) may be replaced by the much simplified demand that

$$h(o) = \left.\frac{da(t)}{dt}\right|_{t=0} = \left.\frac{d|g(t)|^2}{dt}\right|_{t=0}$$

be as large as possible. As shown in Ref. 1, using Krein's theorem and Schwarz's inequality, the impulse-response which is optimum for this case is given by

$$h(t) = \frac{\omega_o}{2\pi}\left[\frac{\sin(\omega_o t/2)}{\omega_o t/2}\right]^2$$

and its frequency-response by the triangular characteristic

$$H(\omega) = 1 - |\omega/\omega_o| \quad \text{for} \quad |\omega| \leq \omega_o$$

$$H(\omega) = 0 \quad \text{for} \quad |\omega| > \omega_o .$$

These responses therefore represent the asymptotic solution for small values of T. It turns out, also, that they provide a servicable approximation for values of $\omega_o T$ less or not much larger than 1. This is illustrated in Figure 3, where computed system functions are displayed for $\omega_o T=2$, and values of the asymptotic solution are indicated by small circles.

The second special case, $T \to \infty$, may be handled[5] by investigating the differential equation satisfied by $\psi_o(c,t)$. Properly scaling the resulting expressions, the Gaussian functions

$$G(\omega) \to k \exp\left(-\frac{T\omega^2}{\omega_o}\right)$$

$$H(\omega) \to \exp\left(-\frac{T\omega^2}{2\omega_o}\right)$$

$$g(t) \to \frac{\sqrt{2}}{k} \exp\left(-\frac{\omega_o t^2}{4T}\right)$$

$$h(t) \to \frac{2}{k^2} \exp\left(-\frac{\omega_o t^2}{2T}\right)$$

are obtained, where

$$k^4 = \frac{8\pi T}{\omega_o}$$

These functions become acceptable approximations (except around t=0 and $|\omega| \geq \omega_o$) for values of $\omega_o T$ much larger than 1. The $\omega_o T = 8$ case is illustrated in Figure 4, with the asymptotic values again indicated by small circles.

IV. FILTER SYNTHESIS

On the basis of the theoretical results obtained so far, the design of a filter approximating the ideal response may be carried out in the following steps:

1. In a typical design situation, the bandwidth ω_o and the desired rise Δa of the step-response a(t) is specified (Figure 2). Δa is usually given in the form x% →(100-x)%; e.g., a given 10% →90% rise represents x=10 and $\Delta a = 1 - 2x/100 = 0.8$. From Δa or x, the parameter c may be found via $\Delta a = 1 - x/50 = \lambda_o(c)$. The fastest possible x% to (100-x)% rise-time T is then given by $T=2c/\omega_o$. A design chart, yielding $\omega_o T$ directly for a specified x, is shown in Figure 5.

2. From ω_o,T and c, the ideal impulse-response $h(t)=[\psi_o(c,t)]^2$ and its transform $H(\omega)$, as well as other transfer functions may be calculated using the theoretical formulas given above.

3. Using existing approximation techniques, realizable transfer characteristics may be found which approach the ideal functions to the desired accuracy. Chebyshev stopband lowpass filters provide a uniform suppression of signals in the range $|\omega| \geq \omega_o$ and can easily be manipulated to approximate a passband loss or time response.[6] Thus, they represent a very convenient tool for carrying out this final step of the synthesis procedure.

To illustrate the design process, an example is next shown. Let the specified values be $\omega_o=1$, T=4. By Figure 5, these values correspond to the minimization of a 21% to 79% rise-time. The response curves for a realization via a seventh-order Chebyshev stopband filter are shown in Figure 6. The filter has a 60 db minimum attenuation for $\omega_o \geq 1$. Its passband response was obtained simply by matching it to the desired value at 7 points in the passband, including the origin. This matching was carried out in the mapped frequency plane (z-plane[6]) and required only the solution of a linear system. The resulting error of the approximation was less than ·025 db for $|\omega| \leq 0{\cdot}8$.

The ideal system function has no phase or linear phase associated with it. The minimum phase actually realized by the Chebyshev stopband filter turns out to be quite linear up to about $|\omega| = 0{\cdot}8$, as illustrated in Figure 6. It was found that phase-equalization affected the time-responses only by insignificant amounts.

A shortcoming of the resulting impulse-response was that, corresponding to the single loss-pole at $\omega \to \infty$, it had a finite nonzero value at t=0. A minor modification of the design process[7] can produce a double loss-pole at $\omega \to \infty$ and thus zero initial value for h(t).

V. EXTENSIONS OF THE THEORY

The results reported here represent, it is hoped, a basis for more comprehensive investigation. Topics which are of interest and seem to be within reach include extension of the technique to bandpass system; taking into account some prescribed nonzero transmission in the "stop-band", as well as finding the final rational system function directly, rather than through the presently used two-stage process.

Of particular interest in future investigations is the controlled relaxation of some specifications, such as absolute bandlimiting and zero ringing and overshoot. To illustrate the latter point, let us assume that the impulse response is permitted to become negative by an amount ϵ:

$$-\varepsilon \leq h(t) < \infty$$

where ε is a specified undershoot, $0 < \varepsilon << 1$. Now it is possible to write

$$h(t) + \varepsilon[k(t)]^2 = |g(t)|^2$$

where $g(t)$, $k(t)$ are bandlimited to $\omega_o/2$ and $\sup k(t) \leq 1$. All previous derivations remain then applicable for $g(t)$. [t]

As a simple and permissible (although not very practical) example, let

$$k(t) = \frac{\sin \omega_1 t}{\omega_1 t} \quad , \omega_1 << \omega_o$$

(Figure 7a) and

$$g(t) = \sqrt{\gamma}\ \psi_o(\omega_o T/2, t), \gamma > 0 .$$

Then

$$h(t) = \gamma[\psi_o(\omega_o T/2, t)]^2 - \varepsilon \sin^2 \omega_1 t/(\omega_1 t)^2 .$$

Integrating from $-\infty$ to ∞,

$$1 = \gamma - \varepsilon\pi/\omega_1$$

is obtained, while integrating from $-T$ to T and utilizing $\omega_1 << \omega_o$,

$$\Delta a \cong \gamma\lambda_o - 2\varepsilon T = \lambda_o + \varepsilon(\lambda_o \pi/\omega_1 - 2T).$$

Denoting by Δa_o the value of Δa for $\varepsilon=0$, we get

$$\frac{\Delta a}{\Delta a_o} \cong 1 + \varepsilon\pi/\omega_1 = 1 + \varepsilon T_\varepsilon$$

where $T_\varepsilon = \pi/\omega_1$ is the approximate duration of the ringing (Figure 7a). The impulse-response itself becomes

$$h(t) = (1+\varepsilon T_\varepsilon)[\psi_o(\omega_o T/2, t)]^2 - \varepsilon \sin^2 \omega_1 t/(\omega_1 t)^2$$

As $\omega_1 \to 0$, the quantities T_ε, Δa and $h(o)$ all become arbitrarily large. The corresponding step-response and frequency-response (Figure 7b) are quite impractical; nevertheless, the results indicate that great gains (of the order $\varepsilon T_\varepsilon$) may be obtained in the speed of the time response by permitting even small undershoots of extended duration.

ACKNOWLEDGMENT

The author is indebted to Mr. Ch. Steele and Dr. M. Gyi for skillful computation of many of the numerical results; also, to Mr. H.J. Orchard and Drs. C. Flammer and J.F. Kaiser for useful discussions.

REFERENCES

1. W.A. Michael, P.A. Szego, G.C. Temes and H.J. Orchard, "An Optimum Bandlimited System with Monotonic Step Response," Proc. IEEE (Corr.) vol. 55, No. 5, May 1967.

2. N.I. Akhieser, Theory of Approximation, New York: Ungar, 1956, p. 154.

3. D. Slepian and H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — Part I," BSTJ, vol. 40, No. 1, Jan. 1961.

4. H.J. Landau and H.O. Pollak, ibid., Part II, or A. Papoulis, The Fourier Integral and Its Applications, New York:. McGraw-Hill, 1962, pp. 72-73.

5. J.-A. Ville and J. Bouzitat, "Note Sur Un Signal de Durée Finie et D'Énergie Filtrée Maximum," Cables & Trans., vol. 11, No. 2, 1957.

6. G.C. Temes and M. Gyi, "Design of Filters With Arbitrary Passband and Chebyshev Stopband Attenuation," 1967 IEEE Inter-nat. Conv. Rec., Pt.5,

7. H.J. Orchard and G.C. Temes, "Filter Design Using Transformed Variables," IEEE Trans. on Circuit Theory, vol. CT-15, No. 4, Dec. 1968.

LIST OF FIGURES

FIGURE 1(a) PERMISSIBLE IMPULSE-RESPONSE

FIGURE 1(b) PERMISSIBLE FREQUENCY-RESPONSE

FIGURE 2 PERMISSIBLE STEP-RESPONSE

FIGURE 3(a) PROLATE SPHEROIDAL FUNCTION FOR $\omega_o T = 2$

FIGURE 3(b) IMPULSE-RESPONSE FOR $\omega_o T = 2$

FIGURE 3(c) FREQUENCY-RESPONSE FOR $\omega_o T = 2$

FIGURE 3(d) STEP-RESPONSE FOR $\omega_o T = 2$

FIGURE 4(a) PROLATE SPHEROIDAL FUNCTION FOR $\omega_o T = 8$

FIGURE 4(b) IMPULSE-RESPONSE FOR $\omega_o T = 8$

FIGURE 4(c) FREQUENCY-RESPONSE FOR $\omega_o T = 8$

FIGURE 4(d) STEP-RESPONSE FOR $\omega_o T = 8$

FIGURE 5 THE SMALLEST $\omega_o T$ AS FUNCTION OF $\underline{x}$

FIGURE 6(a) REALIZED IMPULSE-RESPONSE, $\omega_o T = 4$

FIGURE 6(b) REALIZED STEP-RESPONSE, $\omega_o T = 4$

FIGURE 6(c) REALIZED FREQUENCY-RESPONSE, $\omega_o T = 4$

FIGURE 6(d) REALIZED LOSS-RESPONSE, $\omega_o T = 4$

FIGURE 6(e) REALIZED PHASE-RESPONSE, $\omega_o T = 4$

FIGURE 7(a) IMPULSE-RESPONSE WITH UNDERSHOOT

FIGURE 7(b) FREQUENCY-RESPONSE

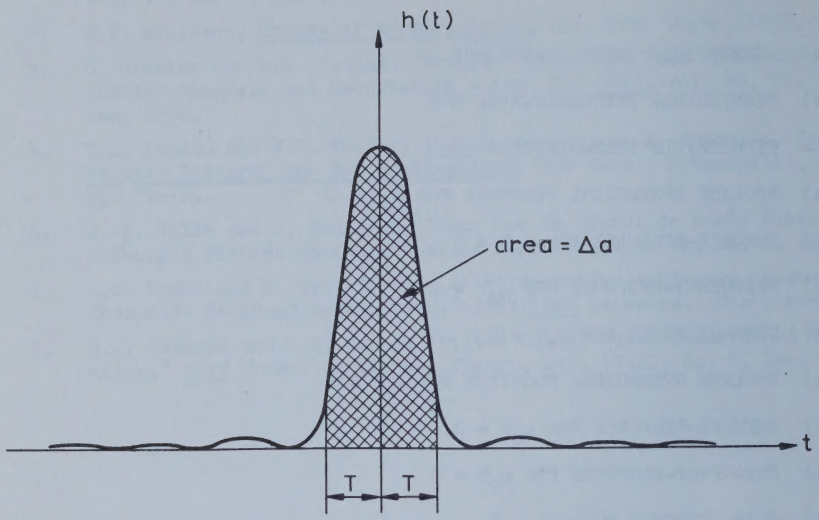


Figure 1(a)

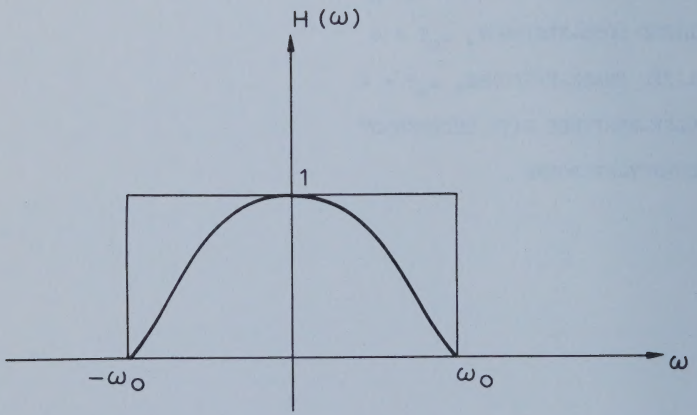


Figure 1(b)

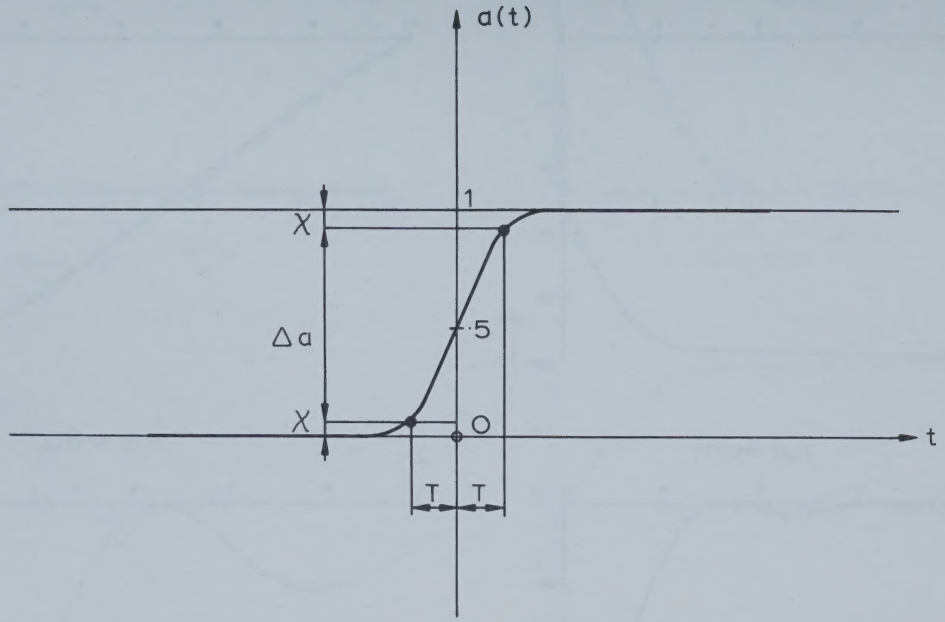


Figure 2

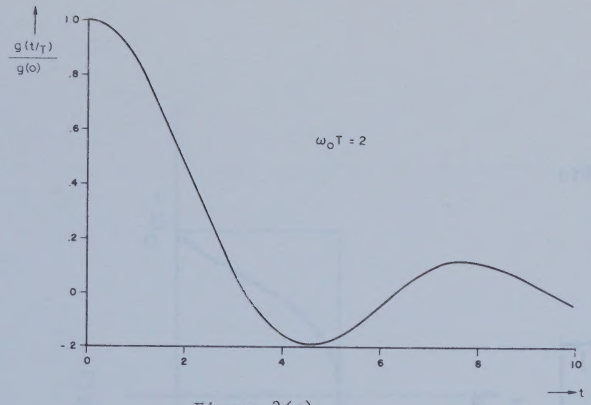


Figure 3(a)

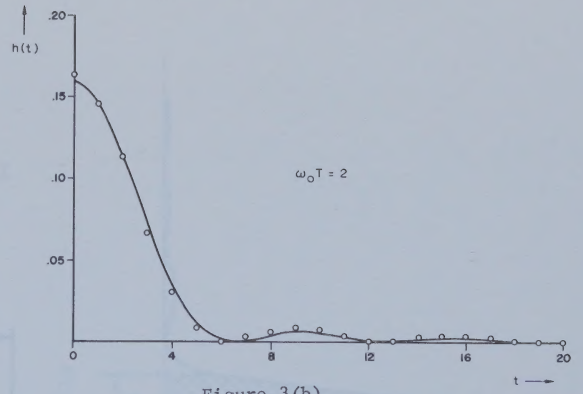


Figure 3(b)

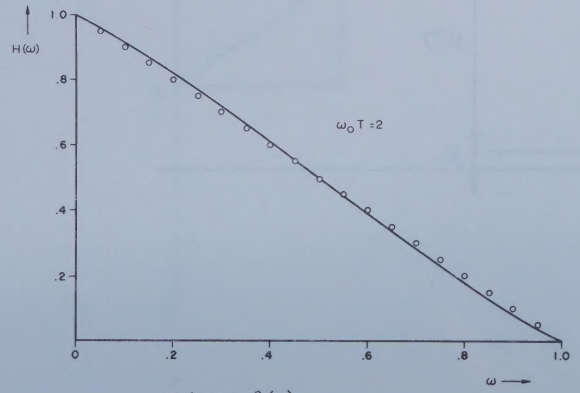


Figure 3(c)

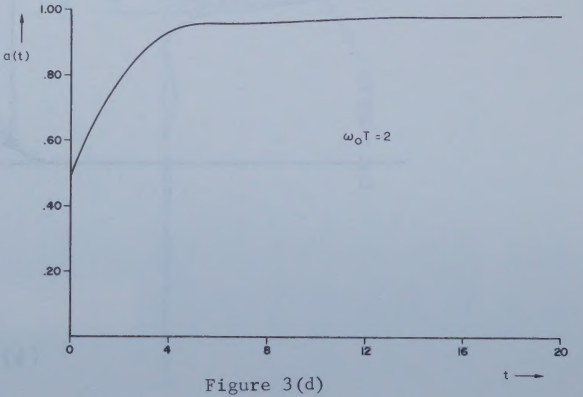


Figure 3(d)

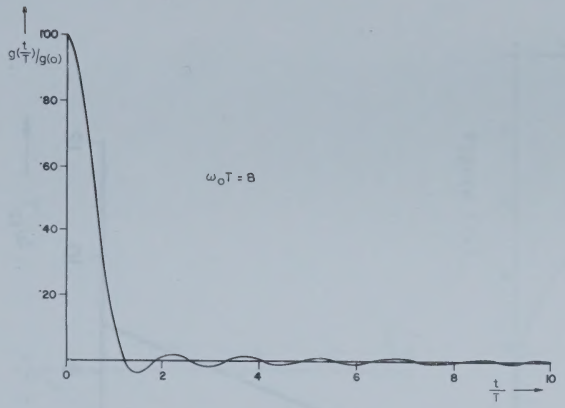


Figure 4(a)

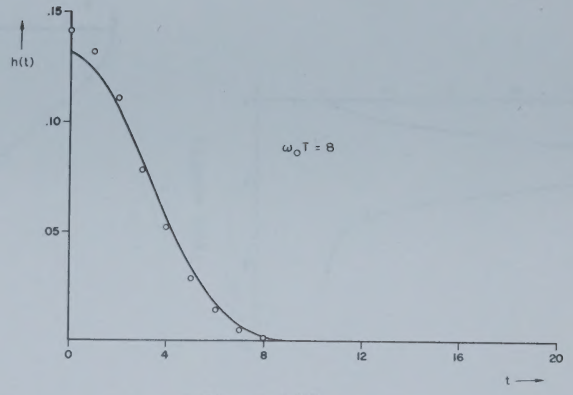


Figure 4(b)

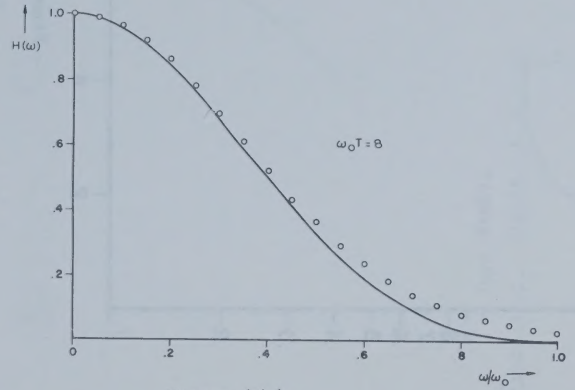


Figure 4(c)

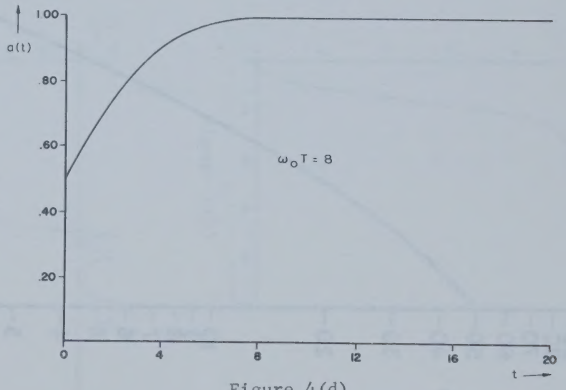


Figure 4(d)

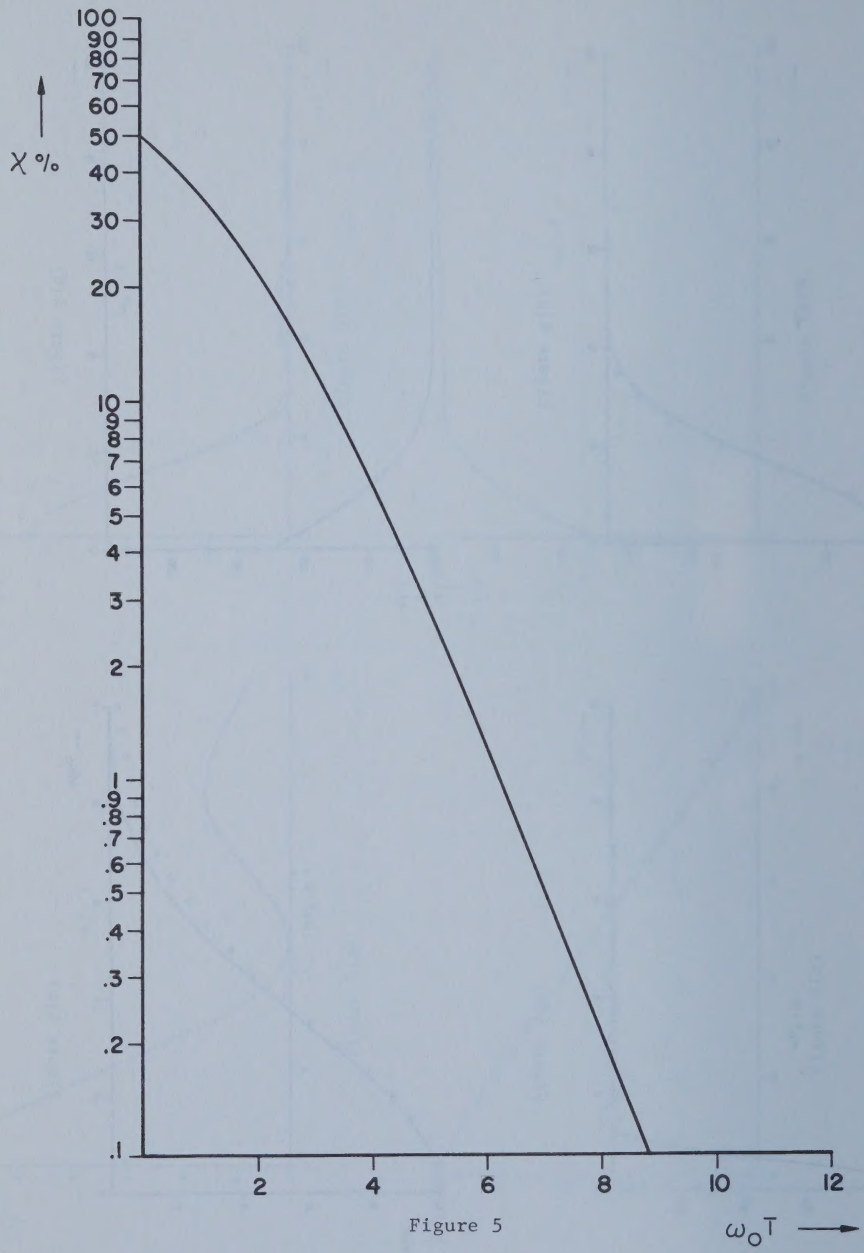

100
90
80
70
60
50
40
30
20
10
9
8
7
6
5
4
3
2
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
X %
2
4
6
8
10
12
$\omega_o T$

Figure 5

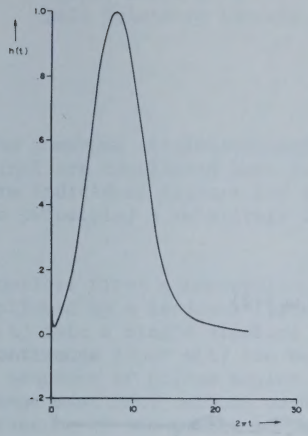


Figure 6(a)

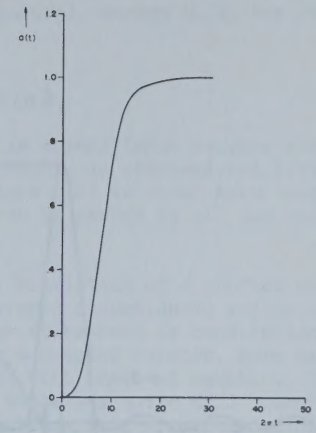


Figure 6(b)

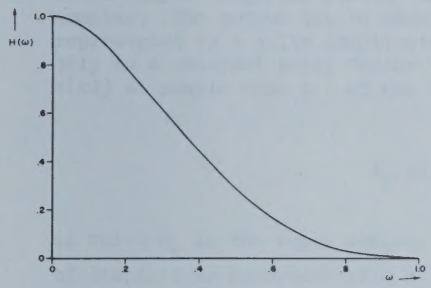


Figure 6(c)

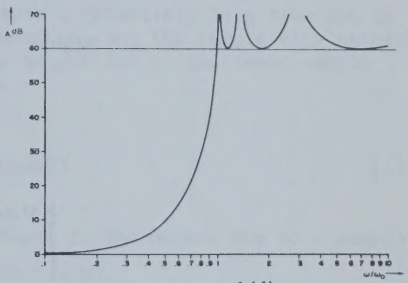


Figure 6(d)

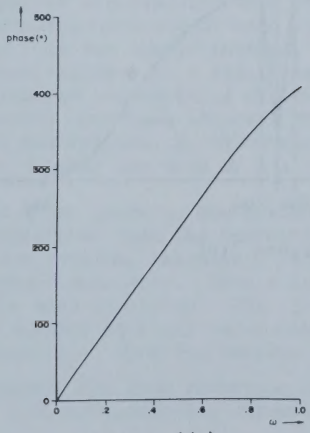


Figure 6(e)

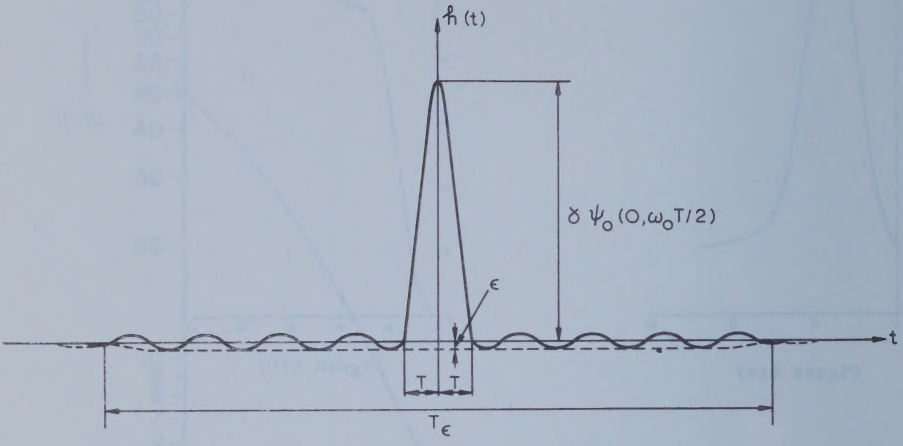


Figure 7(a)

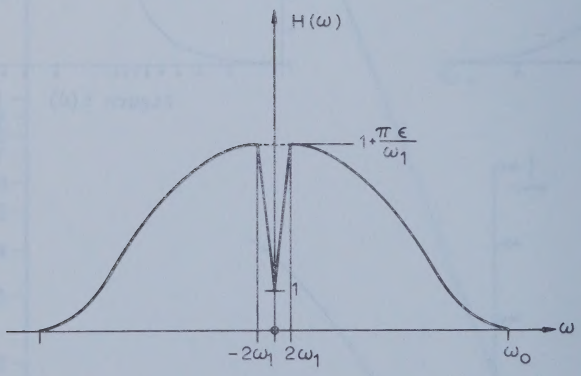


Figure 7(b)

A MODULATION SCHEME FOR LARGE CARRIER SYSTEM TERMINALS

SIDNEY DARLINGTON
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

The numerous single-sideband modulators in a very large carrier system terminal are considered here as a single system. In conventional systems there are individual filters for all the channels. It is shown here that (at least in principle) a relatively few filters can be shared by all the channels.

****

Consider first a conventional modulator, consisting of a product modulator followed by a bandpass filter, which converts a continuous voice-band signal s(t) into a single sideband S(t). If the voice-band is bandlimited, the continuous input s(t) can be replaced by a sampled version, more exactly by a sequence of pulses scaled in accordance with baseband samples. The pulse shape must meet certain conditions, but these are quite mild because of the presence of the filter in the modulator. Fig. 1 is a corresponding block diagram.

Because the modulator is linear in the signal (although time-varying), the output may be regarded as the sum of the outputs due to the individual samples. The output due to each sample lasts a relatively long time and is represented by a quite complicated time function, but the sample size enters only as a constant scale factor. Thus the output due to the input sample s(kT) at sample time t = kT can be written

$$S_k(t) = \frac{H_k}{H_o} S_{ko}(t-kT) \qquad (1)$$

in which H_k is the input sample, and $S_{ko}(t-kT)$ is the output due to a sample of standard or reference size H_o applied at t = kT.

Equation (1) suggests the mechanization described by Fig. 2. The reference pulses $s_o(t-kT)$ are supplied sequentially to a number of different modulators because the single-sideband equivalent of each sample lasts for a number of sample intervals. Thus T is the sample interval and NT is the length of the interval outside of which response to a single sample is negligible. The voiceband samples are clamped sequentially in the N clamp circuits, C, and each clamp circuit is reset once each interval NT. The clamp circuits supply factors to the N multipliers, X, which correspond to the factors H_k/H_o in (1). We shall examine the size of N a little later.

For modulating a single voice channel, the system described by Fig. 2 is obviously much more complicated than the conventional modulator. Suppose, however, a carrier system terminal includes a large number of channels, each requiring a single-sideband modulator. Then a large number of channels can share the same reference time functions. Fig. 3 is a corresponding block diagram. As drawn, it assumes that all voice bands are first modulated onto the same carrier frequency ω_c. Then the samples of all the channels, taken at the same time kT, require the same reference time function $S_o(t-kT)$ and only the sample, clamp and multiplication circuits need be multiplied.

Distribution to other carrier frequencies can be accomplished by further modulations, with only rudimentary filters. Alternatively, the further modulations can be applied to the reference time functions themselves, before the multiplication by observed samples. (In conventional carrier systems, the first modulators use no more than 12 carrier frequencies.)

As an example, suppose that the reference time functions are significant for as long as N = 100 sample intervals, but that there are 10,000 channels to be modulated. Then Fig. 3 requires only 100 filters, compared with 10,000 in a conventional system. On the other hand, the cost of saving 9,900 filters includes 10,000 N or one million relatively simple clamp and multiplication circuits. Presumably large numbers of such devices would be combined in large scale integrated circuits.

Corresponding schemes for demodulation are so similar that we shall note only important differences. The single sideband S(t) corresponding to one of our bandlimited voiceband signals s(t) can be expressed as follows:

$$S(t) = S_c(t)\cos[(\omega_c + \omega_o/2)t] + S_s(t)\sin[(\omega_c + \omega_o/2)t] \qquad (2)$$

in which $\omega_o = 2\pi f_o$, $f_o = 1/(2T)$, and $S_c(t)$ and $S_s(t)$ have bandwidths $< f_o/2$. The functions $S_c(t)$ and $S_s(t)$ can be obtained from product modulations of S(t) with sine and cosine of $(\omega_c + \omega_o/2)t$.

In (2), $S_c(t)$ and $S_s(t)$ can be replaced by sampled versions. Then the demodulated signal s(t) can be expressed as a sum of contributions from the various samples and each contribution will be proportional to a reference time function. Thus

$$s_k(t) = \frac{H_{ck}}{H_o} s_{co}(t-kT) + \frac{H_{sk}}{H_o} s_{so}(t-kT) \qquad (3)$$

in which H_{ck} and H_{sk} are samples of $S_c(t)$ and $S_s(t)$, respectively, H_o is a reference sample, and $s_{co}(t-kT)$ and $s_{so}(t-kT)$ are reference time functions.

An obvious way to generate the reference time functions is to multiply a pulse proportional to H_o by $\cos[(\omega_c + \omega_o/2)t]$ and $\sin[(\omega_c + \omega_o/2)t]$, and then use the two products as the inputs of two conventional single-sideband demodulators. Each demodulator can consist of a bandpass filter, followed by rectification. However, substantially the same result will be obtained more simply if one filter is omitted and the output of the other is multiplied by both $\cos(\omega_c t)$ and $\sin(\omega_c t)$.

In a multichannel system, at least rudimentary prefiltering is also needed to eliminate "foldover" due to the sampling. The foldover can shift frequencies in unwanted channels into the wanted channel. With a carrier frequency spacing f_o and sample rate $1/T = 2f_o$, the channels next above and below the wanted channel will contribute no spurious signal, but the next channels thereafter will do so unless they are eliminated by a filter, which can have bandwidth $< f_o$ and transition intervals $> f_o$. With a sample rate $2/T$, the transition intervals can be $> 3f_o$.

Given the concept of many channels sharing reference time functions, it is easy to devise a very large number of other possible mechanizations. For example, the reference time functions can be generated by reference pulses applied to modified Weaver modulators instead of conventional modulators. (See [1], [2] for the Weaver modulators.) Then standard pulses can be applied to lowpass filters, and the same filters can be used for both modulation and demodulation. Other possibilities include hybrid computers, in which the reference time functions are stored digitally, and analog computers, in which the time functions are stored as a set of precision voltages. (Time functions which differ only by translations of the time scale need be stored only once.)

We have still to examine the number N of sample intervals over which our reference time functions are significantly different from zero. An estimate of the order of magnitude of N can be derived from the frequency functions illustrated in Fig. 4.

In Fig. 4A, $Y_o(\omega)$ is a function of frequency which is real when ω is real and small when $\omega > g\omega_o/2$, $g < 1$. In Fig. 4B, $-Y_o(\omega)$ and $+Y_o(\omega)$ have been translated over frequency intervals $\pm\omega_o/2$. In Fig. 4C, the function in Fig. 4B has been integrated in accordance with

$$Y_{LP}(\omega) = K\int_{-\infty}^{\omega} [Y_o(\lambda + \omega_o/2) - Y_o(\lambda - \omega_o/2)]d\lambda \tag{4}$$

in which K is an arbitrary constant. The result is a lowpass filter characteristic in which the effective bandwidth of $Y_o(\omega)$ determines the width of the transition interval. For a bandpass filter with a symmetrical passband, and transition intervals centered at ω_c and $\omega_c + \omega_o$, one can now use

$$Y_{BP}(\omega) = e^{i\beta}Y_{LP}\left(\omega - \omega_c - \frac{\omega_o}{2}\right) + e^{-i\beta}Y_{LP}\left(\omega + \omega_c + \frac{\omega_o}{2}\right) \tag{5}$$

in which β is an arbitrary constant.

The impulse responses corresponding to the frequency functions (4) and (5) are

$$A_{LP}(t) = KA_o(t)\frac{\sin\frac{\omega_o}{2}t}{t} \tag{6}$$

$$A_{BP}(t) = 2A_{LP}(t)\cos[(\omega_c + \omega_o/2)t+\beta]$$

in which $A_o(t)$ is the transform of $Y_o(\omega)$.

One can now choose functions $Y_o(\omega)$ for which the transforms are available

in analytic form, adjust parameters to meet practical filter requirements, and then determine the effective duration of $A_{LP}(t)$. The author has computed N for functions $Y_o(\omega)$ proportional to $\exp(-a^2\omega^2)$, $\text{sech}^2(a\omega)$, and $\text{sech}(a\omega)$, with a such that $|Y_{LP}(\omega)/Y_{LP}(0)|$ varies by only .001 from 1 and 0, in the pass and attenuation bands respectively, assuming a transition interval from $0.9\omega_o/2$ to $1.1\omega_o/2$. The corresponding impulse response was called negligible when $|A_{LP}(t)| \leq .001\,|A_{LPmax}|$. The N's determined in this way, for the three different functions $Y_o(\omega)$, were 91, 107, and 149.

The filter functions described above have zero phase shift, and adding a linear phase will not change the duration of the impulse responses. The usual filters in analog modulators introduce significant nonlinear phases. This may or may not increase N significantly. It may perhaps even decrease N. Given an arbitrary frequency function, it is not known what phase shift will minimize the corresponding impulse response. However, it is known that the best phase is generally nonlinear. Some phase correction can be added to the filter, if needed.

REFERENCES

1. D. K. Weaver, Jr., "A third method of generation and detection of single-sideband signals", Proc. IRE, vol. 44, pp. 1703-1705, December 1956.

2. S. Darlington, "Some circuits for single-sideband modulation", Proc. of the 3rd Annual Princeton Conference on Information Sciences and Systems, pp. 66-70, March 1969.

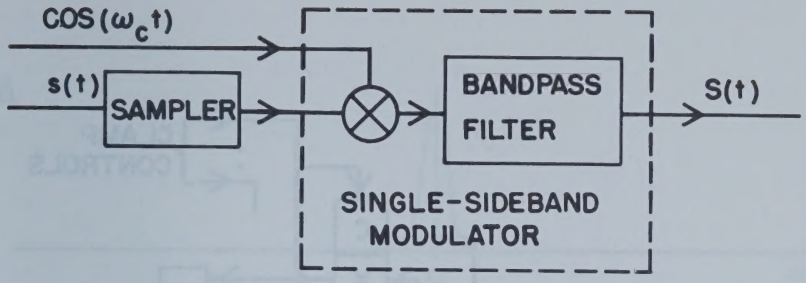


FIG. 1, A MODULATOR WITH SAMPLED INPUT

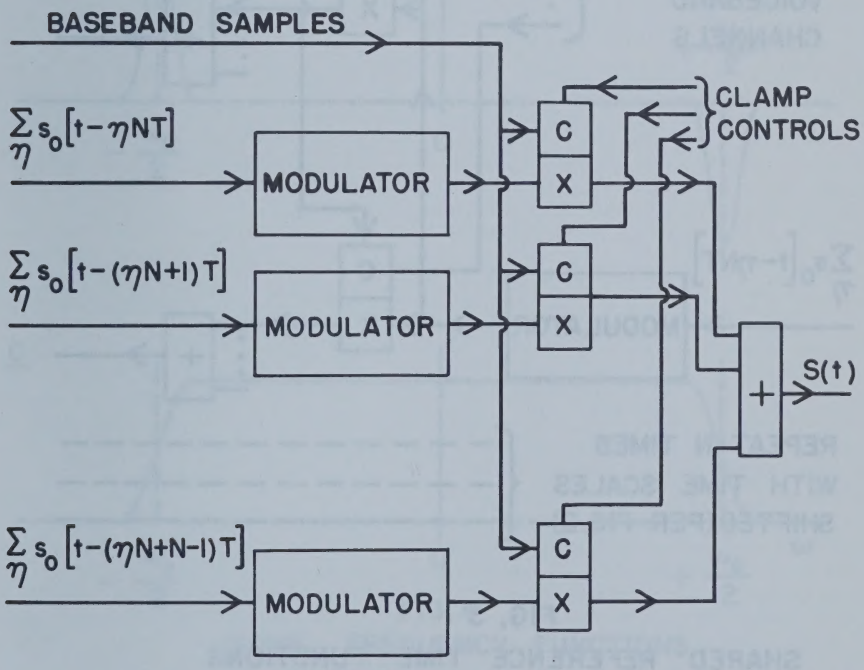


FIG. 2

A MODULATOR USING REFERENCE TIME FUNCTIONS

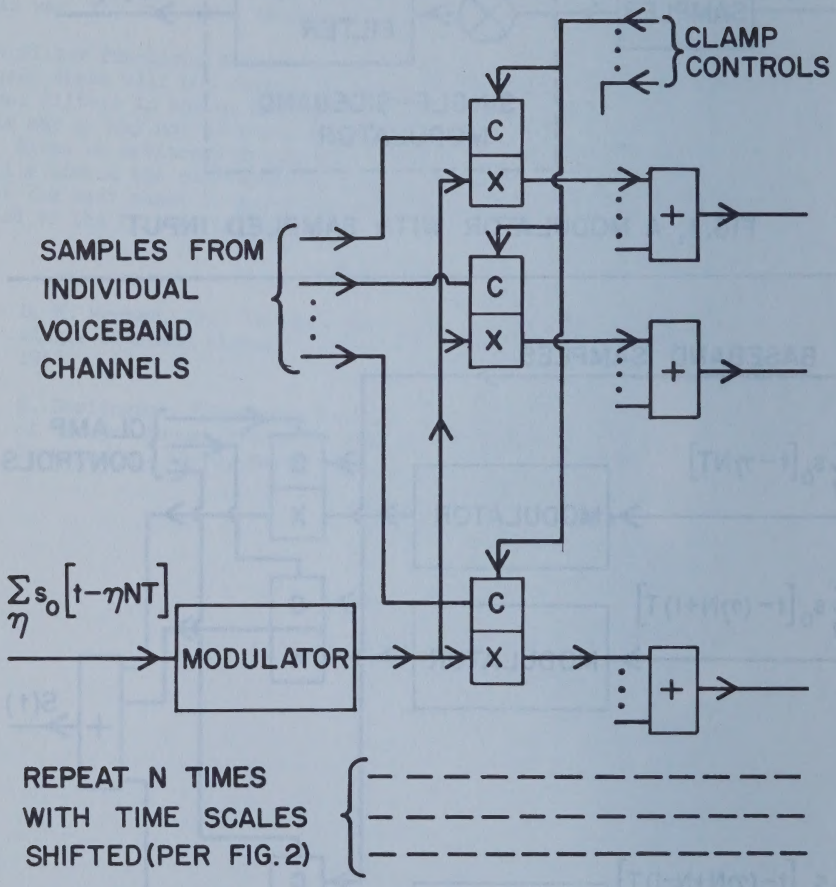


FIG. 3
SHARED REFERENCE TIME FUNCTIONS

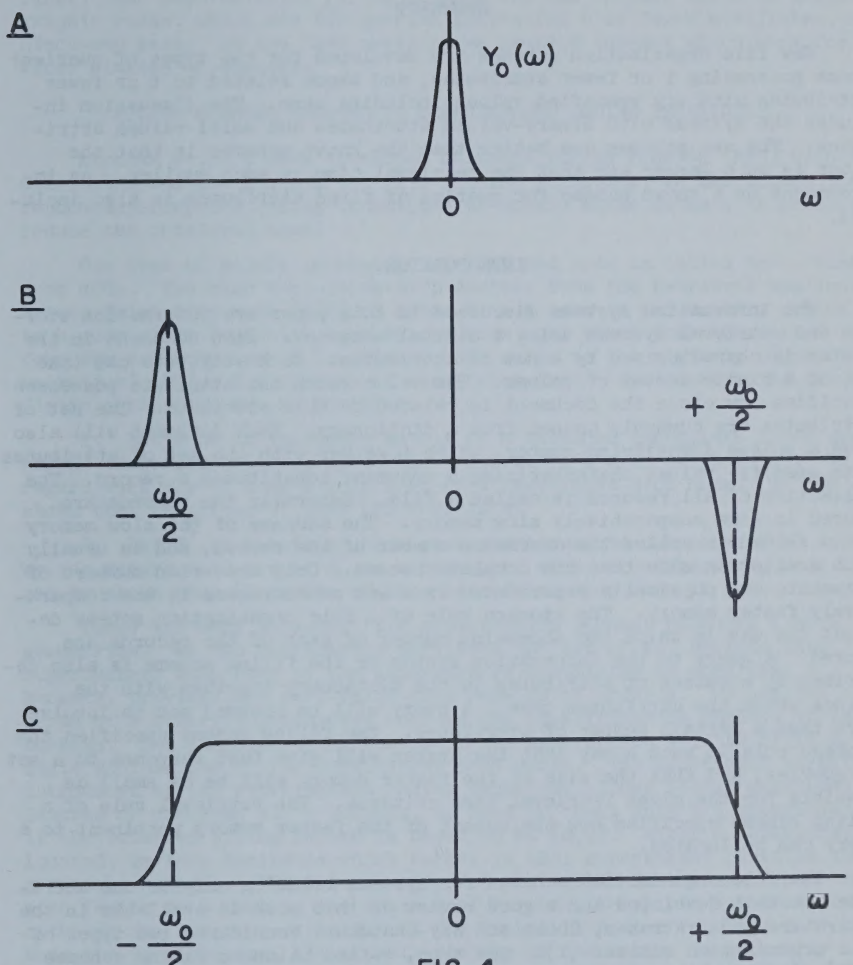


FIG. 4
SOME FREQUENCY FUNCTIONS

SOME RESULTS ON INFORMATION SYSTEMS

DAVID K. CHOW
United Aircraft Research Laboratories
East Hartford, Conn. 06108

ABSTRACT

New file organization schemes are developed for two types of queries: those possessing t or fewer attributes, and those related to t or fewer attributes with any specified values including zero. The discussion includes the systems with binary-valued attributes and multi-valued attributes. The new schemes are better than the known schemes in that the class is much larger and that the retrieval time is much smaller. An improvement on a known scheme for queries of fixed attributes is also included.

INTRODUCTION

The information systems discussed in this paper are information storage and retrieval systems using a digital computer. Each document in the system is characterized by a set of attributes. Each attribute can take one of a finite number of values. The value which the attribute possesses specifies how close the document is related to this attribute. The set of attributes are commonly chosen from a dictionary. Each document will also have a unique identifying number, which together with the set of attributes with specific values characterizing a document constitutes a record. The collection of all records is called a file. Generally the records are stored in some comparatively slow memory. The address of the slow memory for a record is called the accession number of the record, and is usually much smaller in size than the complete record. Only accession numbers of documents are physically represented in a set of addresses in the comparatively faster memory. The storage rule of a file organization scheme defines the way in which the accession number of each of the records are stored. A query to the information system or the filing scheme is also described by a subset of attributes in the dictionary together with the values which the attributes take. A query will be assumed not to involve more than a certain number of attributes. The filing scheme specifies the storage rule in such a way that the system will give fast response to a set of queries, and that the size of the faster memory will be as small as possible for the given retrieval time criteria. The retrieval rule of a filing scheme specifies how the subset of the faster memory pertinent to a query can be located.

The file organization schemes for systems relating only to one attribute are well developed and a good review of this work is available in the literature (4). Abraham, Ghosh and Ray-Chaudhuri considered two types of file organization schemes (1). One type, called balanced filing schemes (BFS), is suitable for a system where queries are related to a fixed number of attributes. The other type, called generalized balanced filing schemes, is suitable for a system where queries possess g attributes for any $g \leq t$. The new balanced filing schemes (NBFS) developed by the author represent a much wider class than the BFS (7). The NBFS also have much faster retrieval response than the BFS and have smaller or comparable storage requirements. Ghosh and Abraham generalized their previous work to files with multiple-valued attributes (8). The discussion in (7) showed that variations of the NBFS are generally better than the previously developed schemes (1), (8) in the class size, retrieval time and storage requirements.

Ray-Chaudhuri developed filing schemes with binary-valued attributes using combinatorial mathematics (11). Other pertinent work related to this paper is based on algebraic coding theory (3), (6).

This paper is concerned with finding better schemes or extending and improving known schemes for information storage and retrieval systems. Filing schemes related to superimposed coding techniques will be discussed first. The combinatorial filing schemes and the schemes based on lexicographic ranks, which are for queries possessing t or fewer attributes, are discussed next. In the last section, we present schemes which work for wider classes of queries.

FILING SCHEMES RELATED TO SUPERIMPOSED CODING TECHNIQUES

Abraham et.al., have developed balanced filing schemes (BFS) with k attributes for k = 2 and k = 3 (1). In the following, we use binary nonrandom superimposed coding techniques to modify these schemes in order to reduce the retrieval time.

One type of binary nonrandom superimposed code is called zero-false-drop code. The name zero-false-drop derives from the retrieval application. A zero-false-drop code (ZFD) of order h is a code in which every sum of up to h code words logically includes no code word other than those used to form the sum (the sum of two binary code words is the digit-by-digit OR operation) (10). Two classes of zero-false-drop codes constructed from finite geometrics are as follows:

1. Let S be an incidence matrix of a balanced incomplete block design (BIBD) whose rows correspond to the u-spaces ($1 \leq u \leq m-1$) of PG (m,q). We can regard each row of S as a code word of a zero-false-drop code. It can be shown that the order of this ZFD code is q.

2. Let S' be a matrix whose rows correspond to u-spaces of EG (m,q). We can regard each row of S' as a code word of a zero-false-drop code. It can be shown that the order of this zero-false-drop code is q-1.

There are many balanced file organization schemes based on finite geometries. We can arbitrarily partition the set of all buckets into superbuckets with each of the superbuckets containing no more than h buckets, provided the zero-false-drop code is of order h. We can use code words of a binary superimposed code as the identification numbers of the buckets. The identification number of the superbucket will be digit-by-digit Boolean sum of the constituent code words. The additional structure introduced by the superimposed coding technique allows faster retrieval. Instead of checking all b buckets, we need to check only b/q superbuckets if the balanced filing scheme is based on PG (m,q). After a superbucket is located, we then determine which bucket in this superbucket contains the pertinent records. The determination of the bucket in the new scheme is much faster than that in the schemes of Abraham, Ghosh and Ray-Chaudhuri. Although the new representation of the bucket ID number requires more bits than the previous one, the overall storage of our scheme is generally less because the computer program to perform the computation is smaller.

Example 1. Let S be an incidence matrix of a balanced incomplete block design. The rows of S are (1110000), (1001100), (0101010), (1000011), (0100101), (0011001) and (0010110) respectively.

Each column position of S corresponds to a point in PG (2,2), and each row of S corresponds to a line in the PG (2,2). We take the row vectors to represent buckets corresponding to the lines of PG (2,2). The representation of a bucket in this scheme uses 7-bits, while the representation of a bucket uses 3-bits in the balanced filing scheme. The new

representation allows much faster retrieval. A ZFD code of order 2 exists with the row vectors of S as code words. Let the row i correspond to a bucket B_i. We can have four superbuckets C_1, C_2, C_3 and C_4 each of which contains the buckets (B_1, B_2), (B_3, B_4), (B_5, B_6) and (B_7) respectively. The ID numbers of C_1, C_2, C_3 and C_4 are (1111100), (1101011), (0111101) and (0010110) respectively.

Let τ be the time required to match one machine word with another. Assume that the time required to test whether a machine word logically includes another is τ'. If the bucket is in C_j ($j = 1,2$ and 3), the number of representations of the superbuckets that need to be tested is j. If a bucket is not the first bucket in C_j ($1 \leq j \leq 3$), it must be the second bucket of C_j. Hence the number of buckets that need to be tested in C_j is one. The last bucket, B_7, can be determined after the ID numbers of the first three superbuckets have been tested. Thus the average time required to determine which bucket contains records pertinent to a query is:

$$T = [(\tau'+\tau)+(\tau'+\tau)+(2\tau'+\tau)+(2\tau'+\tau)+(3\tau'+\tau)+(3\tau'+\tau)+3\tau']/7 = [(15\tau'+6\tau)]/7.$$

To compare two words A and B involves the following operations: (1) Clear Add A, (2) Subtract B and (3) Transfer on zero. The sequence of operations in the following is one way to test if the word A covers the word B: (1) Clear Add $\overline{A}$, (2) AND B and (3) Transfer on zero. The word $\overline{A}$ can be stored prior to the operations. Thus we can assume that τ' is equal to τ. Thus: $T = 3\tau$.

Example 2. Let the number of attributes $n = 31$, and $k = 2$. A BFS_2 can be constructed from PG (2,5). An incidence matrix of a BIBD can be formed with its row vectors corresponding to the lines of PG (2,5). There exists a ZFD code of order 5 with the row vectors as code words. There are 31 buckets. We can group the buckets into superbuckets as follows: The first six superbuckets contain five buckets and the last superbucket contains only one bucket. The average time required to determine which bucket is pertinent to a query is:

$$T = [5(1+2+3+4+5+6)+6]\tau'/31+6(1+2+3+4+4)\tau/31 = 3.58\tau'+2.71\tau$$

Assume $\tau = \tau'$; then $T = 6.29\tau$. If no superbucket is used, the average time required to determine which bucket is pertinent to a query is found to be 16τ.

COMBINATORIAL FILING SCHEMES

Ray-Chaudhuri (11) has developed two classes of single stage filing schemes as follows: one based on the combinatorial configuration with each block corresponding to a t'-dimensional space of PG (r,q) for $r>t'$, and the other based on a configuration constructed from finite geometries. As the determination of the number of buckets in these two classes is difficult in most cases, the designer of information storage and retrieval systems cannot evaluate these schemes easily. In the following, we develop several classes of filing schemes with parameters explicitly known. The combinatorial configurations on which the new filing schemes are based come from several sources. (2), (9), (10).

Filing Schemes Derived from Perfect Codes

Let ℓ be the code length of a binary perfect code and d be the minimum distance of this code. Let b be the number of code words whose weight is d.

Let d = 2e+1. We can construct a filing scheme with each bit position of the ℓ-bit code word corresponding to an attribute. Each of the b code words with minimum weight corresponds to a bucket. Each bucket is associated with d attributes. For any query word with e+1 attributes, this query is covered by one and only one code word corresponding to a bucket. Any vector of weight $\leq$ e is covered by some vector of weight e+1, hence it is covered by some code vector with minimum weight. The storage rule of Ray-Chaudhuri's filing schemes can be applied to the schemes derived from perfect codes.

Example 3. A Golay (23,12) code is a perfect code with minimum distance d = 7. A filing scheme with 23 attributes can be made. Each bucket is associated with 7 attributes. The number of code words with weight 7 is 253. Hence the number of buckets is 253.

Tactical configurations can be made from nonbinary perfect codes also (2). Filing schemes can be constructed from these tactical configurations in a similar way, as the binary case.

Filing Schemes Derived From Steiner Systems

A d-th order Steiner System can be defined as follows: Let S be a set of ℓ elements, and let D_i be a collection of nonempty i-sets of S, i = 3, 4, ... , d. Define a free t-set of S as one not containing any member of $D_3 \cup \ldots \cup D_t$ for t = 2,3,4, ..., d-1, let every free t-set be contained in precisely one member of D_{t+1} and every proper subset of every element of D_{t+1} be free. Then $\{D_i\}_{i=3}^{d}$ is a d-th order Steiner System. A filing scheme can be constructed from a Steiner Triple System (d=3). Let ℓ be the number of required attributes, let d be the number of attributes associated with any one bucket, and let b be the number of buckets. The filing schemes constructed are for queries with two attributes or less. This class exists for the parameters (d,ℓ,b) satisfying the condition. $\ell = 1$ or 3 (mod 6), d = 3 and $b = \ell(\ell-1)/6$.

A filing scheme for queries with four attributes or less can be constructed from the d-th order Steiner System for $d \geq 4$. The number of attributes in this filing scheme is $\ell = 2^{d-1}$. A query of weight 2 is pertinent to one and only element in D_3. Thus the number of buckets in D_3 is $b_3 = \binom{\ell}{2}/\binom{3}{2} = \binom{\ell}{2}/3$. D_4 is a collection of 4-sets of S such that every 3-set of S, not a member of D_3, is contained in precisely one element of D_4. Thus the number of buckets in D_4 is $b_4 = [\binom{\ell}{3} - b_3]/\binom{4}{3}$. For example, when d = 4, ℓ = 7, b_3 = 7 and b_4 = 7. The total number of buckets b for a filing scheme based on a fourth order Steiner System is 14.

Filing Schemes Derived From Certain Block Codes

The block codes we are concerned with here are the codes which are invariant under t-fold transitive permutation groups. A permutation group G on the symbols $a_1, a_2, \ldots a_\ell$ which contains some permutation p replacing any whatever given ordered set of t symbols by any whatever other given ordered set of t symbols is called t-fold transitive. We can denote the bit positions of a block code of length ℓ as $a_1, a_2, \ldots, a_\ell$. If after applying the permutation p to a block code C the resulting set is also C, C is invariant under G. Let B be the collection of all code words of C

with weight d. Let $d \geq t$. For any ℓ-tuple U of weight t, there exists an element V in B such that V covers U. Thus we can construct a filing scheme for queries of weight less or equal to t. Each bucket corresponds to an element of B. The number of buckets is equal to the number of code words in C with weight equal to d. Let B' be a subset of C such that every member of B' has weight equal to d. If B' is invariant under G, we can construct a filing scheme based on B'. The class of filing schemes based on B' includes the filing schemes based on B as a special case. However, the special case is much easier to construct because of our knowledge of the weight distributions of the codes. For example, reference (9) contains weight distributions for some BCH codes which are invariant under a doubly transitive permutation group. New balanced incomplete block designs can be obtained from a knowledge of the weight distributions of the codes.

SCHEMES BASED ON LEXICOGRAPHIC RANKS AND THEIR PARTITION

The schemes based on lexicographic ranks allow their parameters to take almost any values. Thus they allow information system designers much more freedom.

Let the filing schemes consisting of ℓ attributes A_i $(0 \leq i < \ell)$ and b buckets B_j $(0 \leq j < b)$. We represent the attribute A_i by an integer i and the buckets B_j by an integer j. Chow (1969) established new balanced filing schemes (NBFS) for queries of k attributes where k is a fixed integer (7). Each bucket of an NBFS contains records pertinent to c sets of k attributes where c is related to the parameters ℓ, k and b. For a query consisting of k attributes A_{j_1}, A_{j_2},--- and A_{j_k}, the computation of the bucket containing the pertinent records uses the number $j_1 j_2 \ldots j_k$ where $j_1 < j_2 \ldots < j_k$ (7). We can modify the NBFS so that the schemes can be used to store and retrieve documents pertinent to the query of g attributes $g \leq k$. If a query consists of g attributes A_{i_1}, A_{i_2}, ... A_{i_g} where $i_1 < i_2 < \ldots < i_g$, a mapping which we shall define later, will carry $i_1 i_2 \ldots i_g$ to $j_1 j_2 \ldots j_k$ where $\{j_1, j_2, \ldots, j_k\}$ contains $\{i_1, i_2, \ldots i_g\}$ as a subset. All documents pertinent to $A_{i_1}, A_{i_2}, \ldots A_{i_g}$ will be stored in the bucket which contains all documents pertinent to $A_{j_1}, A_{j_2}, \ldots, A_{j_k}$. Thus the computation method for determining the bucket pertinent to $j_1 j_2 \ldots j_k$ can be used. In the following we shall illustrate the mapping of $i_1 i_2 \ldots i_g$ to $j_1 j_2 \ldots j_k$ through examples for k = 2 and 3. Extention to any k can be made in similar manner. The method of partitioning the documents in a bucket into subbuckets in reference (11) can be used for our scheme.

Example 4. Let $(k, \ell, b) = (2,7,7)$. The seven buckets B_i $(0 \leq i \leq 6)$ of an NBFS are (01,02,03), (04,05,06), (12,13,14), (15,16,23), (24,25,26), (34,35,36), and (45,46,56) respectively. If the element ij is in bucket B_v, all the documents pertinent to a query with attributes A_i and A_j are in B_v.

The rule of mapping i_1 to $j_1 j_2$ is as follows: If $i_1 \leq \ell-2=5$, i_1 is mapped to $i_1(i_1+1)$. If $i_1 = \ell-1$, i_1 is mapped to $(\ell-2)(\ell-1)$. Thus the seven buckets of the new filing schemes are (0,01,02,03), (04,05,06), (1, 12,13,14), (2,15,16,23), (24,25,26), (3,34,35,36), (4,5,6,45,46,56).

Example 5. Let k = 3. Both $i_1 i_2$ and i_1 are mapped $(\ell-3)(\ell-2)(\ell-1)$ when $i_1 \geq \ell-3$. When $i_1 < \ell-3$, i_1 is mapped to $i_1(i_1+1)(i_1+2)$. When $i_1 < \ell-3$ and $i_2 = i_1+1$, $i_1 i_2$ is mapped to $i_1(i_1+1)(i_1+2)$. When $i_1 < \ell-3$ and $i_2 \neq i_1+1$, $i_1 i_2$ is mapped into $i_1(i_1+1)i_2$.

The method used in analyzing the storage requirements of the NBFS can be used in analyzing the storage requirements of the new scheme. We assume uniformly distributed data, all M records appear once and only once where $M = 2^{\ell}$. The storage requirement for the new scheme with parameters $(k,\ell,b) = (2,7,7)$ is found to be 3.93M. We have evaluated the storage requirement for Ray-Chaudhuri's scheme with the same parameters and found it to be 4.37M.

EXTENSION OF PREVIOUS SCHEMES

The new information systems discussed previously are capable of efficient retrieval of additional queries. Let us consider the following example.

Example 6. Let $(d,\ell,b,t) = (3,7,7,2)$. This scheme has the same parameters as that of Example 1 of reference (11). This filing scheme is based on PG (2,2) which has seven lines. The seven lines consist of points (1,2,4), (1,3,5), (2,3,6), (1,6,7), (2,5,7), (3,4,7) and (4,5,6), respectively. One bucket is associated with each line. This scheme was designed to retrieve all queries with two or less attributes. There are seven subbuckets ($1\ \bar{2}\ \bar{4}$, $\bar{1}\ 2\ \bar{4}$, $\bar{1}\ \bar{2}\ 4$, $1\ 2\ \bar{4}$, $1\ \bar{2}\ 4$, $\bar{1}\ 2\ 4$, $1\ 2\ 4$). The subbucket $i\bar{j}k$ contains all documents which have attributes A_i and A_k and does not have the attribute A_j. We can use this scheme for additional queries. Suppose we want all documents pertinent to attributes A_1 and not pertinent to attributes A_2 and A_4. In other words, we want all documents pertinent to queries $(A_1\bar{A}_2\bar{A}_4)$ or $(1\ \bar{2}\ \bar{4})$. Obviously, we can obtain these documents easily because they are in a subbucket of a bucket. Thus the documents pertinent to queries ($1\ \bar{2}\ \bar{4}$, $\bar{1}\ 2\ \bar{4}$, $\bar{1}\ \bar{2}\ 4$, $1\ 2\ \bar{4}$, $1\ \bar{2}\ 4$, $\bar{1}\ 2\ 4$, $1\ 2\ 4$) can be retrieved efficiently from the first bucket. We can also obtain all documents pertinent to each of the following queries ($1\ \bar{2}$, $1\ \bar{4}$, $2\ \bar{4}$, $\bar{1}\ 2$, $\bar{1}\ 4$, $\bar{2}\ 4$, $1\ 2\ \bar{4}$, $1\ \bar{2}\ 4$, $\bar{1}\ 2\ 4$). The scheme is originally used for retrieval of all documents pertinent to queries (1, 12, 14, 24). The use of this scheme for additional queries is no more difficult than the original usage. This filing scheme cannot retrieve queries ($\bar{1}$, $\bar{2}$, $\bar{4}$, $\bar{1}\ \bar{2}$, $\bar{1}\ \bar{4}$, $\bar{2}\ \bar{4}$, $\bar{1}\ \bar{2}\ \bar{4}$) since we cannot obtain all pertinent documents from the first bucket alone. An efficient way to test whether the documents pertinent to a query can be retrieved easily is to test whether all product terms of the query are covered by the subbuckets. For example, the query (12) can be written as the sum of two product terms (124, $12\bar{4}$). Both of (124, $12\bar{4}$) are in the first bucket. Hence all documents pertinent to the query (12) can be retrieved efficiently. The query ($\bar{1}$) can be written as the sum of four product terms ($\bar{1}\ \bar{2}\ \bar{4}$, $\bar{1}\ \bar{2}\ 4$, $\bar{1}\ 2\ \bar{4}$, $\bar{1}\ 2\ 4$). Since ($\bar{1}\ \bar{2}\ \bar{4}$) is not in the first bucket, the query ($\bar{1}$) cannot be retrieved efficiently.

Bose, et.al., developed filing schemes based on algebraic coding theory (3). Their schemes are for queries related to t or fewer attributes with any specified values. The query Q has attributes $A_{j_1}, A_{j_2}, \ldots, A_{j_g}$ and each A_{j_i} has a specified value V_{j_i} where for schemes based on a q-ary code ($g \leq t$), V_{j_i} takes one of the q values 0,1, ..., q-1. A record

for which the attributes $A_{j_1}, A_{j_2}, \ldots, A_{j_g}$ have the values $V_{j_1}, V_{j_2}, \ldots$ V_{j_g} is said to satisfy the query, or pertain to the query. Their scheme allows the following additional freedom: When $q = 2$, V_{j_1} takes two values 0,1. We can change the notation of the query by introducing the compliment of A_j. $\bar{A}_{j_i}$ means $V_{j_i} = 0$ and A_{j_i} means $V_{j_i} = 1$. Thus a query can be written as a product of g variables, each variable being either the attribute A_j or its compliment $\bar{A}_j$. The query of the form $A_{j_1}\bar{A}_{j_2}$ is not allowed in Ray-Chaudhuri's scheme.

In the beginning of this section, we discussed some additional usage of the storage schemes of Ray-Chaudhuri. In Example 6, the filing schemes are for the query having two or less attributes. It cannot retrieve all documents which are not pertinent to attribute A_i, or pertinent to $\bar{A}_i$. With additional storage, we can extend the Ray-Chaudhuri scheme for the type of queries discussed in Bose, et.al., (1967). Take the same parameters of Example 6. The bucket associated with points (i,j,k) in the new scheme will consist of eight subbuckets ($\bar{i}\,\bar{j}\,\bar{k}$, $\bar{i}\,\bar{j}\,k$, $\bar{i}\,j\,\bar{k}$, $\bar{i}_{\ell}j\,k$, $i\,\bar{j}\,\bar{k}$, $i\,\bar{j}\,k$, $i\,j\,\bar{k}$, $i\,j\,k$). For uniformly distributed data, $M = 2^{\ell}$. Each bucket will have all M records. Total storage requirement is $bM = 7M$.

The following example gives some comparisons between the new scheme and that of Bose, et.al., (1967).

Consider a filing scheme based on the binary Hamming Code of length 7 and constructed according to Bose et.al., (1967). The minimum distance of this code is 3; therefore, the filing scheme is for queries related to 2 or fewer attributes with any specified values. The scheme has 7 buckets. Each bucket consists of $2^7 - 1$ subbuckets. For uniformly distributed data, each bucket contains $M - 1 = 2^7 - 1$ records. The storage requirement for this scheme is $7(M - 1)$. In this example, all allowable queries are distributed very unevenly in relation with the buckets. The new scheme can be constructed from PG (2,2) with a bucket associated with the PG (2,2) itself, which bucket contains all M records. Thus the new scheme requires about one-seventh of the storage requirement of the scheme based on coding theory. Since $d = \ell = 7$ in the new scheme, all queries which are related to 7 or fewer attributes with any specified value can be retrieved efficiently.

The new schemes we have considered thus far are related to binary-valued attributes. The previous schemes can be generalized for multi-valued attributes. A record is said to be pertinent to the attributes $A_{i_1}, A_{i_2}, \ldots, A_{i_g}$ with each of them taking values V_{i_j} $(1 \leq j \leq g)$ if the record possesses the same values V_{i_j} for these A_{i_j}. The storage and retrieval rule of schemes with multiple-valued attributes can be devised from the binary-valued attribute scheme as follows: We defined a mapping which takes V_{i_j} to 1 for nonzero V_{i_j} and take V_{i_j} to 0 if V_{i_j} is zero. The distribution of documents in the buckets will be treated in the same way as in the binary scheme after we have performed the mapping. We then partition the subbucket into second stage subbuckets according to the V_{i_j}. For example, if a subbucket related to $(\bar{u}\bar{v}w)$ contains all documents with

$V_{i_v} = 0$, $V_{i_u} \neq 0$ and $V_{i_w} \neq 0$. Two documents are in the same second stage subbucket if and only if they have the same V_{i_u} and V_{i_w}.

CONCLUSIONS

With the new presentation of the bucket identification number of the balanced filing schemes and the use of superimposed coding techniques, the modified balanced filing scheme will have faster retrieval response than the known balanced filing scheme. The time saving is generally largest when the number of buckets is approximately equal to the number of super-buckets.

The combinatorial filing schemes developed include many classes of schemes whose parameters are explicitly known. The schemes based on lexicographic ranks and their partition allow their parameters to take almost any values. Thus they provide much more freedom for information system designers. Since computations in higher dimensional finite fields are time consuming, the schemes derived from new balanced filing schemes generally have faster responses than the schemes based on finite geometrics or algebraic coding theory.

The known schemes based on block codes allow query attributes to take any specified value. We have developed new schemes for the same type of queries which have larger class size and smaller retrieval time.

ACKNOWLEDGMENT

The author is indebted to Dr. J. H. Bennett for many valuable discussions during the preparation of this paper.

REFERENCES

1. Abraham, C. T., Ghosh, S. P. and Ray-Chaudhuri, D. K. (1968), "File Organization Schemes Based on Finite Geometries," Information and Control, 12, 143-163.

2. Assmus, E. F., Jr., Mattson, H. F., Jr. and Turyn, R. (1966) "Cyclic Codes" AFCRL-66-348.

3. Bose, R. C., Abraham, C. T. and Ghosh, S. P. (1967), "File Organization of Records for Multiple Valued Attributes for Multiattribute Queries," IBM Report RC-1886.

4. Buchholz, W. (1963), "File Organization and Addressing," IBM System Journal, 2, 86-111.

5. Carmichael, R. D. (1937), "Introduction to the Theory of Groups of Finite Order," Dover Publications, Inc., New York, N.Y.

6. Chow, D. K. (1967), "A Geometric Approach to Coding Theory With Application to Information Retrieval," Report R-368, Coordinated Science Laboratory, University of Illinois.

7. Chow, D. K. (1969), "New Balanced File Organization Schemes," To appear in Information and Control.

8. Ghosh, S. P. and Abraham, C. T. (1968), "Application of Finite Geometry in File Organization for Records with Multiple-Valued Attributes," IBM Journal of Research and Development, 12, No. 2, 180-187.

9. Kasami, T. (1966). "Weight Distributions of Bose-Chaudhuri-Hocquenghem Codes," Report R-317, Coordinated Science Laboratory, University of Illinois.

10. Kautz, W. H. and Singleton, R. C. (1964), "Nonrandom Binary Superimposed Codes," IEEE Trans., IT-10, No. 4, 363-378.

11. Ray-Chaudhuri, D. K. (1968), "Combinatorial Information Retrieval Systems for Files," SIAM, Journal of Applied Mathematics, 16, No. 5, 973-992.

A NEW DEPARTURE IN INFORMATION PROCESSING BY FACSIMILE COMMUNICATION

IRA KAMEN, Senior Member IEEE
Vice President, Comfax Communications Corp., New York, N. Y.
CHARLES McDEVITT
President, ComputerPix Corp., New York, N. Y.

The ability to retrieve documents by computer techniques in milliseconds and not to be able to disseminate them by high speed low cost communication methods has handicapped the application of information retrieval systems in their use by business, industry and medicine.

The system to be described is both a technical and economical breakthrough. This system of transmitting retrieved documents stored in microform by low cost error-free, all electronic facsimile means should make facsimile communications turn a corner offering opportunities for many new computer applications.

This system, a joint development of the Comfax Communications Corp. and ComputerPix Corp., will make it possible for professionals of all types to retrieve documents from stored research libraries at the rate of one minute per document employing ordinary telephone lines.

The importance of the one minute rate is that the telephone company on a nationwide basis has filed tariffs applicable in 1970 which will make the cost of a phone call using one minute facsimile equipment less than the charges currently applied to special delivery letters.

This development, coupled with the fact that all information is being placed into microform, will make computer oriented information utilities in the 70's as common as other utility services are now.

Until the Comfax/ComputerPix system, a COM library did only half the job. Once a microform was retrieved it had to be put into an on-site viewer, the desired information selected manually and then read. Or, the selected material had to be enlarged, printed and mailed to the inquirer.

A slow process which defeated the expedition of the computer retrieval process and which was subject to clerical error, required a human interface, and could suffer the vagaries of the mails.

Even using the best facsimile systems, picture quality was limited and transmission--and line charges--were 6 minutes a page. The systems were also subject to faults and maintenance inherent in mechanical operations.

Today, the Comfax/ComputerPix system provides new dimensions to the distribution of information. Comfax/ComputerPix system permits the executive to dial up the information with total automation. Instead of having to manipulate a viewer or wait for the mails, he is presented with an 8 1/2 x 11" hard copy--over ordinary dial-up telephone lines--by the ComputerPix receiver in less than 50 seconds. The accelerated speed is accomplished by variable velocity scanning involving a unique scanner which in effect "skips" white space, transmitting only opaque, informational material. It is so fast that the average signature can be verified in less than 20 seconds.

Obviously, the utility of the Comfax/ComputerPix system is virtually universal. It applies to any field in which reference to archives is

required. In fact, some 80 to 95% of industry classifications keep records stored on microfilm; their problem is rapid low cost access to them--away from the site of storage. The Comfax/ComputerPix system adds impetus to the burgeoning communications industry ($14,100,014,000 in 1968 and currently growing at the rate of 8-10% per year). Data transmission alone, growing at 20% per year, will account for some $10 billion in volume by 1975, paralleling the computer industry's growth (current volume--$9,822,000,000).

But Comfax/ComputerPix applications will have their greatest impact on the information industry--the cataloging of archival material and expenditures made in its retrieval for use. The information industry generated a volume of $500,000,000 in 1968. Projections for 1969 are at $1,380,000,000. This will grow to $3,780,000,000 by 1971.

Facsimile equipments currently available on the market such as the Xerox Telecopier II, the Magnafax 850 and the Graphic Sciences unit provide an 8 1/2 x 11" copy of approximately 96 line per inch density (equivalent to approximately 1,000 TV lines) in six minutes over the dial-up network, with the full telephone charges of AT&T for more than 6 minutes. Rapid retrieval of information with the above times is not considered to be acceptable and although facsimile units are available which decrease this time, they also require high cost broader band channels of communication. The digital coding techniques that others are employing to reach the design goals described for the Comfax system result in a terminal cost beyond the reach of most information retrieval systems. Scanning the information on the document, storing it in the memory, processing the information to select the most efficient algorithm and, in digital form, sending the information through a 4800 bit modem onto the telephone line are the characteristics of these systems. The Comfax/ComputerPix system using variable velocity scanning does not require the use of a special modem and transmits the signal in high speed analog form over the regular telephone line. This results in a considerably lower cost.

The ComputerPix method of converting retrieved documents in microform to the facsimile mode to be described has many physical variations depending upon whether it is adapted to the Ampex, IBM, Mosler, Sanders, Recordak, or other information retrieval equipment associated with dial access computers.

Before describing the specific facsimile development we should first understand the principles of the variable velocity scanning system, which was conceived for narrow bandwidth TV.

Narrow band TV implies a method of transmission of images within a frequency bandwidth much narrower than is now needed for the transmission of images processed by the conventional TV principles. Narrow band TV is not only needed, but the need for narrow bandwidth TV is becoming acute. Transmission of images has become an essential facility interwoven with the rest of our technological development. As in any other case of original abundance, we may have been wasteful with the natural resources. Our

conventional TV system is using sufficient frequency bandwidth capable of transmitting a checkerboard-like picture changing from white to black during every cell and even reversing black to white during every frame. Is this requirement which we have imposed on our TV system realistic? If it is, it is only in theory; it is doubtful that out of the astronomical number of frames transmitted since TV's birth, one single TV field ever demanded this requirement. Actually the complexity of the average TV picture requires only 1/20 of the available bandwidth. Another factor of 5 is wasted in repeating transmission of identical information in consecutive frames.

The factor of waste in frequency bandwidth is therefore about 100, and pictures using 4.5 mc of frequency bandwidth could probably be transmitted with only 45 kc. By means of variable velocity scanning and careful filming, entire shows could be transmitted with as narrow a bandwidth as is now used by FM radio stations, transmitting high fidelity audio background music over subcarrier channels.

The wastefulness of the present TV system in frequency bandwidth lies in the apriori assumption that a TV system must scan at a uniform rate. The required bandwidth is then automatically fixed by the rise time involved in changing the intensity from white to black to black to white within the time interval the scanning beam takes to move a distance of one horizontal cell. The frequency bandwidth thus becomes independent of the number of intensity changes per image frame.

The paradox of the logic around our conventional TV system may be demonstrated if we were to try to apply same reasoning to the design of an automobile. If we were to assume that the car had to maintain a particular uniform speed, our main design effort would be concentrated around the design of the springs of the car to be capable of withstanding the shock which the springs would receive when the car would go over the deepest road depression. The number of road depressions in this case is an irrelevant matter, and the car will take the same time to complete a trip whether the road has one or a million depressions. It may be noted, however, that if the road were not continuously grooved in the form of a sine or a square wave, a car having accelerator and brakes would complete a trip over the same distance in a shorter time by accelerating over the smooth stretches of the road and slowing down while passing over a road depression. The condition of special car springs is eliminated since the variable speed car can slow down to a speed compatible with whatever springs it has.

This popular exposition example besides demonstrating the absurdity of uniform velocity, also serves to introduce another concept in TV scanning involving variable scanning velocity, which has also been called the Sonic Vee technique. In a scanning system such as this, the electronic beam would scan at velocities higher than the conventional TV over regions of the image where the intensity changes to a speed compatible with the frequency bandwidth of the transmission circuit, connecting transmitter and receiver. The transmission circuit may be a telephone line, an FM radio

subcarrier or even a magnetic tape.

U. S. Patent No. 3,204,026 has been issued for equipping the imaginary car with a driver, the provision of advanced information to help the receiver to slow down in advance of an intensity variation. It is this principle and the sophistications of U. S. Patent Nos. 3,384,710 and 3,459,886, and other patents pending which allows, including the pictorial sorting of microfilm, for one minute facsimile communications of average documents over ordinary dial-up telephone lines using high speed one minute transmissions.

The ComputerPix transmitter and ComputerPix receiver are so designed that microfilm which is retrieved can be converted to hard copy over dial-up telephone lines within one minute.

Reduced to elementals, the system consists of a ComputerPix transmitter within a computer-oriented mass storage device and a hook-up to available telephone broadcast lines. It works this way: 1) A request is received from the computer terminal or by telephone; 2) On signal, the appropriate microform is found in the mass storage device -- in the Mosler 410 storage and sorting system this might be one fiche card out of 200,000; 3) The transmitter scans the appropriate fiche and presents the particular page of information and transmits it to the requester. The operation takes roughly 60 seconds, 10 seconds for the computer to pictorially sort and retrieve and 50 seconds to transmit a document from microform to hard copy over an ordinary telephone line.

The actual type of microform does not matter. It can be an aperture card, a cartridge of film, or a roll of film, so long as it is stored within a mass storage device capable of isolating the desired item. ComputerPix has or will develop the needed interfaces for automatic transmission.

The ComputerPix service can best be seen in a typical, though hypothetical, example. A far west daily newspaper editor, for example, requires all of the clippings that a Chicago newspaper has on Julie Nixon because she will shortly be in the far west city to address a youth rally. Previously, the local paper would have telexed the Chicago paper which keeps its clipping file on microfiche within a Mosler 410 system. The Chicago paper's computer would locate abstracts of all articles and these would be telexed back to the far west editor. He then would determine which articles he wanted and telex a reply to the paper whose computer, interfaced with the 410, would then signal the 410 which fiche were needed. At this point either a duplicate diazo card would be made, or a print of the entire card or of the specific pages. This would be mailed to the local paper.

The ComputerPix service eliminates these last two (2) steps. As soon as the computer signals the 410 and the appropriate fiche are located, the transmitter scans the card, reads it at 125 lines per inch on CRT, and transmits it over telephone lines directly to the far west paper. Therefore, the information is available in one minute rather than one or two days.

The creation of the useable ComputerPix system and the data presented in the following figures and tables are results of the effort of Malcolm Reader, ComputerPix's Vice President.

To best understand the practical system we must review the quantitative and qualitative objectives (corresponding to the above operational objectives) summarized by Figure 1.

1. In 8 1/2" wide format, to print all graphical information which appears in sizes down to .060 letter height; standard typewritten material to be of high quality. This corresponds to 125 lines per inch definition, with finer granularity.

2. To reproduce all signatures and graphical material, exact position of format as is commonly done on high quality five minute facsimile units, in less than one minute on dial-up telephone systems.

3. To accept all microfilm of 8 1/2 x 11" copy down to 30 to 1 reduction of full size. This requires electro-optical modulation transfer functions of 70% at 75 line pairs per millimeter throughout the format.

4. To provide access to microfilm reductions of drawings up to 17 x 22" in microfilm formats. This requires optical options for varying scales of microfilm reduction down to 40 to 1.

The basic technique employed by the development is variable speed scanning as previously described in general terms. The multiple redundancies that exist in letters and other graphical material are well-known and much discussed in engineering literature. Methods for achieving a reduction in bandwidth and/or transmission time to achieve an economy by the elimination of these redundancies have been the object of continuing research. The equipment developed jointly by Comfax and ComputerPix Corps. exploits several aspects of redundancy and adopts the resulting economy to specific cost break-points in readily available communications facilities. In typical 8 1/2 x 11" business written material, ease of understanding by the human observer requires that material be separated by areas, assembled in columns and suitably positioned for ease of understanding. This results in many spaces left unoccupied in the physical format of page type information and its counterpart in microfilm. The same can be said of drawings in that the rules for graphical representation and location of notes require that many large and unoccupied areas appear in the format of the drawing. The development equipments are based on the principle that unoccupied areas will be found and coded so that transmission of information and printing of the received copy is accelerated by rapidly moving through unoccupied areas, then slowed down to information content limited speeds for those areas of copy which contain data. Copy to be transmitted is scanned and converted to a stream of data signals by electro-optics; the stream of data is examined for used and unused areas; then transmitted at a proper rate. Appropriate code commands are inserted to indicate speed of transmission. The receiving station receives this data, separates codes and information and appropriately scans at fast or slow speeds to reconstitute the infor-

mation as originally viewed at the transmitting end. Electro-optics are different for microfilm scanning but electronics for standard and microfilm transmitters are almost identical.

Identical receivers are used in both systems. Nominal definition is 125 lines per inch. However, special analog and digital techniques are employed which provide granularity to .002 in the horizontal direction. The technique is illustrated by two samples to be discussed below.

The first example, Figure 2, is an abstract of a technical paper. The analysis of speed of transmission of the contents of this document appears in Figure 3. The areas contained therein are classified as types I thru V. Each area has its known transmission characteristics and its known speed-up potential due to the appearance of physical unused areas of the copy. Table 1 tabulates the area content by each classification of this document. Figure 4 illustrates the speed-up potential for various run lengths of unused areas and applies these to the probability experience of one classification of area. In Figure 4, Curve 1 illustrates probability distribution of type II areas. This is normalized for area by Curve 2 of the Figure. Curve 3 indicates the speed-up potential for each of the individual run lengths. Curve 4 shows the effective area of type II areas under speed-up. The probability distribution shown in Figure 4 is derived from a statistical distribution of many samples. Quantitative speed from comparision of Curves 2 and 4 shows that on the average acceleration of 1 to .0909 is obtained for type II areas. By a similar analysis, each of the types of areas illustrated in Figure 3 has been determined to permit speed-up of the factors as illustrated in Table 2.

The illustrated copy of Figure 2 with the area breakdown of Table 1 permits total transmission time in 51.4 seconds as tabulated in Table 2.

A schematic and logic diagram combination which is representative of microfilm of drawing data is illustrated in Figure 5. The various elements of the drawing can be analyzed for classification into area types which have frequency of distribution related to corresponding area types of letter text. The breakdown into areas is shown in Table 3. The speed-up factors for types of areas differ in probability distributions from letter format but can be closely approximated from Curve 3 of Figure 4. The results of these probability distributions are listed in Table 4.

Application of these factors by the various quantities of areas by appropriate types results in a total transmission time of 51.0 seconds.

Equipment embodying these principles has been demonstrated and sample transmissions have closely verified the achievement of the speeds derived from analysis shown above. The advantages of decreased horizontal granularity were apparent in copy produced.

In a society where knowledge is being created faster than it can be acquired this development fills an important need.

1. 125 Line (62.5 Line Pairs) Vertically
2. 125 Line Definition with 500 Line Granularity Horizontally
3. Less than 1 minute on Dial-Up Phone Lines. For typical (8% black) copy - 8½" x 11"
4. Compatibility of Facsimile & Microfilm Networks
5. Micro form down to 40:1 reduction
6. Drawing & letter material in mixed microfilm library

Quantitative & Qualitative Achievements

Figure 1

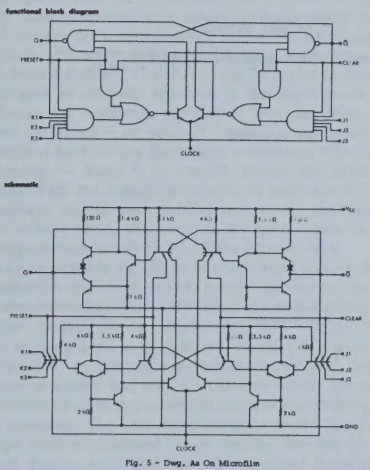


Fig. 5 - Dwg. As On Microfilm

Type I	10.8 sq. in.	11.5%
Type II	9.9 sq. in.	10.6%
Type III	.34 sq. in.	.4%
Type IV	24.9 sq. in.	26.7%
Type V	47.6 sq. in.	50.8%

Table 1 Contents of Fig 2 By Area Type

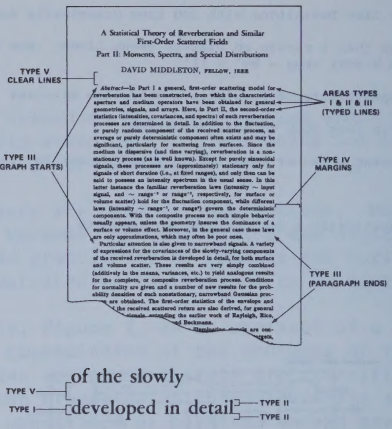

A Statistical Theory of Reverberation and Similar First-Order Scattered Fields
Part II: Moments, Spectra, and Special Distributions
DAVID MIDDLETON, FELLOW, IEEE
TYPE V CLEAR LINES
AREAS TYPES I & II & III (TYPED LINES)
TYPE III (PARAGRAPH STARTS)
TYPE IV MARGINS
TYPE III (PARAGRAPH ENDS)
of the slowly
developed in detail
TYPE V
TYPE I
TYPE II
TYPE II

Fig. 3 Area Classifications

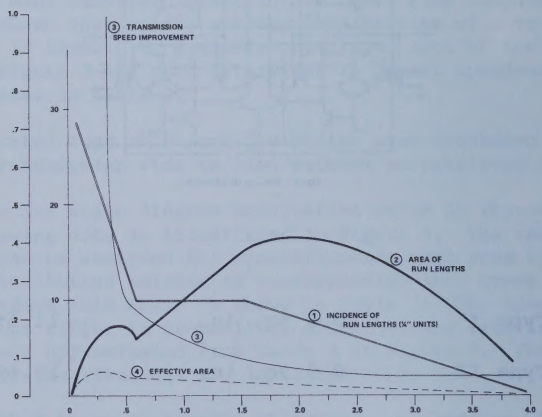


Fig. 4 Analysis of Type II Areas For Speed-Up Ratio

A Statistical Theory of Reverberation and Similar First-Order Scattered Fields

Part II: Moments, Spectra, and Special Distributions

DAVID MIDDLETON, FELLOW, IEEE

Abstract—In Part I a general, first-order scattering model for reverberation has been constructed, from which the characteristic aperture and medium operators have been obtained for general geometries, signals, and arrays. Here, in Part II, the second-order statistics (intensities, covariances, and spectra) of such reverberation processes are determined in detail. In addition to the fluctuation, or purely random component of the received scatter process, an average or purely deterministic component often exists and may be significant, particularly for scattering from surfaces. Since the medium is dispersive (and time varying), reverberation is a nonstationary process (as is well known). Except for purely sinusoidal signals, these processes are (approximately) stationary only for signals of short duration (i.e., at fixed ranges), and only then can be said to possess an intensity spectrum in the usual sense. In this latter instance the familiar reverberation laws (intensity $\sim$ input signal, and $\sim$ range^{-3} or range^{-2}, respectively, for surface or volume scatter) hold for the fluctuation component, while different laws (intensity $\sim$ range^{-2}, or range^{0}) govern the deterministic components. With the composite process no such simple behavior usually appears, unless the geometry insures the dominance of a surface or volume effect. Moreover, in the general case these laws are only approximations, which may often be poor ones.

Particular attention is also given to narrowband signals. A variety of expressions for the covariances of the slowly-varying components of the received reverberation is developed in detail, for both surface and volume scatter. These results are very simply combined (additively in the means, variances, etc.) to yield analogous results for the complete, or composite reverberation process. Conditions for normality are given and a number of new results for the probability densities of such nonstationary, narrowband Gaussian processes are obtained. The first-order statistics of the envelope and phase of the received scattered return are also derived, for general narrowband signals, extending the earlier work of Rayleigh, Rice, Hoyt, Nakagami, Middleton, and Beckmann.

Many special cases for arbitrary illuminating signals are considered, including an approach to complex (distributed) targets, uniform beams, multiple receivers, transmitter and receiver at the same location, large and small doppler, etc., where the general formulation (as in Part I) is broadband, given in terms of frequency-selective apertures, time-varying dynamic cross sections, and arbitrary input signal waveforms.

A number of second-order statistical properties of the medium response is defined and evaluated, which are needed for the combined theoretical and experimental exploration of the medium itself. It is found that the present model of the dispersive reverberation channel cannot be fully described by the statistical structures postulated earlier (and successfully) in radar astronomy: a considerably more involved second-order statistic is required (and obtained) here. To what extent it may provide an adequate description of reality remains to be studied. Part II concludes with a short summary of the principal results and assumptions.

Fig. 2 - Microfilm Frame

Area	(Area) The Reduction Factors		
Type I	11.5% x 1.0	=	11.50%
Type II	10.6% x .0909	=	.96%
Type III	.4% x .075		.03%
Type IV	26.7% .063		1.74%
Type V	50.8% .031		1.58%
	Effective Area		15.81%

15.8% of 300 sec = 47.4 sec

Housekeeping 4.0 sec

51.4 sec

Table 2 Weighted effective area

Type I	Top	4.16	
	Bottom	6.53 in	
		10.69	11.42%
Type II (None)			
Type III (None)			
Type IV	Top	6.75	
	Bottom	13.50	
		20.25	21.65%
Type V		62.5	66.93%

Table 3 Contents of Fig 4 By Area Type

	Area	(Area) The Reduction Factors)		
Type I	11.42%	x1.0	=	11.42%
Type II	None			
Type III	None			
Type IV	21.65%	x0.10		2.17%
Type V	66.93%	x0.031	=	2.06 %
			Effective Area	15.65%

15.65% of 300 sec = 47.0 sec

Housekeeping 4.0 sec

51.0 sec

Table 4 Weighted Effective Area

SENSITIVITY STUDIES OF ANALOG COMPUTER SIMULATIONS

HSUEH-HSIANG CHI AND CHI-TSONG CHEN
Department of Electrical Sciences
State University of New York at Stony Brook
Stony Brook, New York 11790

ABSTRACT

This paper studies the eigenvalue sensitivities of various analog computer simulations of linear time-invariant differential equations and linear time-invariant dynamical equations. The sensitivity of the eigenvalue λ_k due to the variations of all the elements a_{ij} of a matrix is defined as $\sum_{i,j} \left| \frac{\Delta\lambda_k/\lambda_k}{\Delta a_{ij}/a_{ij}} \right|$. It is shown that the Jordan form simulation is less sensitive than the normal form simulation for any differential equation or any dynamical equation which may have repeated real and/or complex eigenvalues. For the differential equation or dynamical equation which has only simple real eigenvalues, it is shown that the Jordan form simulation is less sensitive to the variations of parameter than any other form simulation.

I. INTRODUCTION

This paper studies the eigenvalue sensitivities of analog computer simulations of linear time-invariant differential equations and linear time-invariant dynamical equations. Recently this problem has been given considerable attention [1-9]. In [1-4],were studied the shifts of the characteristic roots or the responses of a specific simulation due to either the perturbations of gains or the imperfactions of operational amplifier. In [5-9] various simulations of the same equation were studied and compared. This paper studies the second problem, namely comparing various simulations of the same equation. It extends the existing results in two aspects: First, there is no restriction except stability imposed on the equations studied, second, a conclusion is established in the most general setting.

The class of differential equations and dynamical equation studied in this paper are assumed to be linear and time-invariant. There is no restriction imposed on their eigenvalues; they may be real or comples; simple or repeated. However it is assumed that all the eigenvalues have negative real parts. This assumption is necessary for practical simulation of a differential equation or a dynamical equation in an analog computer. Without it, the response of the differential equation will tend to infinity and the amplifiers will be saturated.

There are infinity many ways to simulate a differential equation and a dynamical equation on an analog computer. Among these various simulations, it is of theoretical and practical interests to compare these different

simulations. There are many possible criteria to make comparisons: easiness to set up, simplicity in wiring, the number of components used, sensitivity of the behavior of the solution due to the variations of components. Since the eigenvalues (the roots of the characteristic polynomial) of a differential equation or a dynamical equation have the most eminent effect on the solution, the eigenvalue sensitivities will be studied.

This paper is organized as follows: In Section 2 we discuss the set up of analog computer simulations. In Section 3 the eigenvalue sensitivity is defined. In Section 4 we compare the eigenvalue sensitivities of the normal form and the Jordan form simulations of differential equations with repeated real and/or complex eigenvalues. The eigenvalue sensitivity of the normal form and an arbitrary form simulation of differential equations with simple real eigenvalues are studied in Section 5.

II. SET UP OF ANALOG COMPUTER SIMULATIONS

Given a differential equation

$$D(p)y(t) = N(p)u(t) \tag{1}$$

where $D(p) \triangleq p^n - a_1 p^{n-1} - a_2 p^{n-2} - \ldots - a_n$; $N(p) = b_1 p^{n-1} + b_2 p^{n-2} + \ldots + b_n$ and $p^i = d^i/dt^i$, or equivalently, the transfer function

$$G(s) \triangleq \frac{\hat{y}(s)}{\hat{u}(s)} = \frac{N(s)}{D(s)} , \tag{2}$$

the following dynamical equation

$$\dot{\underset{\sim}{x}} = \underset{\sim}{A}\,\underset{\sim}{x} + \underset{\sim}{b}\,u \tag{3a}$$

$$\underset{\sim}{y} = \underset{\sim}{c}\,\underset{\sim}{x} \tag{3b}$$

where

$$\underset{\sim}{A} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ a_n & a_{n-1} & a_{n-2} & \cdots & a_1 \end{bmatrix} , \quad \underset{\sim}{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{4}$$

$$C = [b_n \quad b_{n-1} \; b_{n-2} \quad \ldots \quad b_1]$$

can be easily set up. This equation is said to be in the normal form. By using this equation, an analog computer simulation can be readily set up. Another common way to set up a simulation is by using the Jordan form dynamical equation. For convenience, it is assumed that N(s)/D(s) can be written as

$$\frac{N(s)}{D(s)} = \frac{N(s)}{(s-\lambda_1)^m(s-\lambda_2)^{\ell}(s-\bar{\lambda}_2)^{\ell}} = \frac{N(s)}{(s-\lambda_1)^m(s^2-d_1 s-d_2)^{\ell}}$$

$$= \frac{e_1}{(s-\lambda_1)^m} + \frac{e_2}{(s-\lambda_1)^{m-1}} + \ldots + \frac{e_m}{(s-\lambda_1)}$$

$$+\frac{e_{11}s+e_{12}}{(s^2-d_1s-d_2)^{\ell}}+\frac{e_{21}s+e_{22}}{(s^2-d_1s-d_2)^{\ell-1}}+\cdots+\frac{e_{\ell 1}s+e_{\ell 2}}{(s^2-d_1s-d_2)}\quad . \tag{5}$$

where $d_1=\lambda_2+\bar{\lambda}_2=2\,\mathrm{Re}\,\lambda_2$, $d_2=\lambda_2\cdot\bar{\lambda}_2=|\lambda_2|^2$. Then it can be easily verified that the dynamical equation

$$J:\ \dot{\underset{\sim}{x}}=\begin{bmatrix}\underset{\sim}{A}_1 & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{A}_2\end{bmatrix}\underset{\sim}{x}+\begin{bmatrix}\underset{\sim}{b}_1\\ \underset{\sim}{b}_2\end{bmatrix}u \tag{6a}$$

$$y=\begin{bmatrix}\underset{\sim}{C}_1 & \underset{\sim}{C}_2\end{bmatrix}\underset{\sim}{n} \tag{6b}$$

where

$$\underset{\sim}{A}_1=\begin{bmatrix}\lambda_1 & 1 & 0 & \cdots & 0\\ 0 & \lambda_1 & 1 & \cdots & 0\\ 0 & 0 & \lambda_1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 1\\ 0 & 0 & 0 & \cdots & \lambda_1\end{bmatrix}\quad (m\times m)$$

$$\underset{\sim}{A}_2=\begin{bmatrix}\underset{\sim}{\Phi} & \underset{\sim}{I}_2 & \underset{\sim}{0} & \cdots & \underset{\sim}{0} & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{\Phi} & \underset{\sim}{I}_2 & \cdots & \underset{\sim}{0} & \underset{\sim}{0}\\ \vdots & \vdots & \vdots & & & \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{\Phi} & \underset{\sim}{I}_2\\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{0} & \underset{\sim}{\Phi}\end{bmatrix}\quad (2\ell\times 2\ell)$$

$$\underset{\sim}{\Phi}=\begin{bmatrix}0 & 1\\ d_1 & d_2\end{bmatrix}\qquad \underset{\sim}{I}_2=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

$$\underset{\sim}{b}_1=[0\ 0\ 0\ \cdots\ 1]'\qquad \underset{\sim}{b}_2=[0\ 0\ 0\ \cdots\ 0\ 1\ 1]$$

$$\underset{\sim}{c}_1=[e_1\ e_2\ e_3\ \cdots\ e_m]\qquad \underset{\sim}{c}_2=[e_{11}\ e_{12}\ e_{21}\ e_{22}\ \cdots\ e_{\ell 1}\ e_{\ell 2}],$$

is a realization of N(s)/D(s). This equation is said to be in the modified Jordan form. If a dynamical equation realization of N(s)/D(s) is neither in the form of (4) nor in the form (b), then it is said to be in an arbitrary form.

Given a linear time-invariant differential equation, many dynamical equation realizations can be set up for it. Consequently we may have many analog computer simulations. The normal form dynamical equation can be written up by inspection, hence this simulation is the easiest one to set up. In order to simulate (1) or (2) in the (modified) Jordan form, we have to first compute the roots, and then carry out the partial fraction expansion. Hence the Jordan form simulation is harder to set up than the normal form. In practice, the arbitrary form simulation is seldom used.

The number of elements which are different from 0 or 1 in the normal form dynamical equation and the Jordan form dynamical equation are both equal to 2n. Hence in the simulations, they will use the same number of components (potentiometers and amplifiers). The elements 0 and 1 are not counted in 2n, because in the simulations they are either open or direct connection. Hence as far as the number of components used in the simulation is concerned, the normal form and the (modified) Jordan form simulations are the same. It is obvious that if we simulate a differential equation by using an arbitrary form dynamical equation, the number of components used is

larger than that in the normal or Jordan form.

III. EIGENVALUE SENSITIVITY

In this section we study the sensitivity of analog computer simulations. Since the eigenvalues of a differential equation or a dynamical equation have the most eminent effects on the solution, the eigenvalue sensitivity due to the variations of components will be studied. The eigenvalues are the roots of the characteristic polynomial of the matrix $\underset{\sim}{A}$, hence the eigenvalue sensitivity depends only on the elements of A and has nothing to do with the elements of $\underset{\sim}{b}$ and $\underset{\sim}{c}$.

The eigenvalue sensitivity of λ_k due to the variation of a_{ij} is defined as

$$S_{a_{ij}}^{\lambda k} = \left| \frac{\Delta\lambda_k/\lambda k}{\Delta a_{ij}/a_{ij}} \right| \tag{8}$$

The sensitivity $S_{a_{ij}}^{\lambda k}$ is thus the ratio of the percentage change of λ_k to the percentage change of a_{ij} which causes the change of λ_k. Another possible way to define eigenvalue sensitivity is to use $|\Delta\lambda_k/\Delta a_{ij}|$. However the definition in (8) is used in this paper for the reason that it includes a weighting factor. That is, for the same amount of variation $\Delta\lambda_k$, the sensitivity is larger if the eigenvalue is closer to the origin.

Let us define the following two polynomials

$$f(\lambda, a_{ij}) \triangleq \det(\lambda \underset{\sim}{I} - \underset{\sim}{A}) \tag{9}$$

and

$$f(\lambda, a_{ij}+\Delta a_{ij}) \triangleq \det(\lambda \underset{\sim}{I} - (\underset{\sim}{A}+\Delta\underset{\sim}{A})) \tag{10}$$

where a_{ij} is an element of A; Δa_{ij} is the variation from the nominal value a_{ij}. If λ_k is an eigenvalue of $\underset{\sim}{A}$, that is, a root of $f(\lambda, a_{ij})$, then $f(\lambda_k, a_{ij})=0$. Now if for some reason, the matrix $\underset{\sim}{A}$ changes to $\underset{\sim}{A}+\Delta\underset{\sim}{A}$, then λ_k will not remain to be an eigenvalue of $\underset{\sim}{A}+\Delta\underset{\sim}{A}$. Let $\lambda_k+\Delta\lambda_k$ be an eigenvalue of $(\underset{\sim}{A}+\Delta\underset{\sim}{A})$, that is $f(\lambda_k+\Delta\lambda_k, \Delta a_{ij}+ a_{ij})=0$. Then following the fact that a root of a polynomial is a continuous function of its coefficients, we conclude that if Δa_{ij} is very small, so is $\Delta\lambda_k$. Assume that all the elements of $\underset{\sim}{A}$ except a_{ij} are fixed and expand $f(\lambda_k+\Delta\lambda_k, a_{ij}+\Delta a_{ij})$ into

$$f(\lambda_k+\Delta\lambda_k, a_{ij}+\Delta a_{ij}) = f(\lambda_k, a_{ij}) + \left.\frac{\partial f}{\partial \lambda}\right|_{\lambda_k, a_{ij}} \Delta\lambda k + \left.\frac{\partial f}{\partial a_{ij}}\right|_{\lambda_k, a_{ij}} \Delta a_{ij} + \cdots$$

$$+ \frac{1}{m!}\left(\Delta\lambda_k \frac{\partial}{\partial \lambda} + a_{ij}\frac{\partial}{\partial a_{ij}}\right)^m f \Big|_{\lambda_k, a_{ij}} + \cdots \tag{11}$$

If $\left.\frac{\partial f}{\partial \lambda_k}\right|_{\lambda_k, a_{ij}} \neq 0$ and $\left.\frac{\partial f}{\partial a_{ij}}\right|_{\lambda_k, a_{ij}} \neq 0$, then by neglecting the higher power

of Δa_{ij} and $\Delta\lambda_k$, we obtain

$$S^{\lambda_k}_{a_{ij}} = \left|\frac{\Delta\lambda_k/\lambda_k}{\Delta a_{ij}/a_{ij}}\right| \doteq \left|\frac{a_{ij}\,\partial f/\partial a_{ij}}{\lambda_k\,\partial f/\partial\lambda_k}\right|_{\lambda_k, a_{ij}} \tag{12}$$

This approximation is commonly used in sensitivity study, and it holds only if λ_k is a simple root of $f(\lambda, a_{ij})$. If $\partial f/\partial\lambda_k\big|_{\lambda_k, a_{ij}} = 0$ as it occurs when λ_k is a multiple root of $f(\lambda, a_{ij})$, then equation (12) cannot be used and a more general form has to be employed. Suppose $f(\lambda, a_{ij})$ has root λ_k with multiplicity m, then it is clear that $\partial^{\ell} f(\lambda, a_{ij})/\partial^{\ell}\lambda\big|_{\lambda_k} = 0$, for $\ell = 0, 1, \ldots, m-1$

If $f(\lambda, a_{ij})$ is linear in a_{ij}, then $\dfrac{\partial^{\ell+s} f(\lambda, a_{ij})}{\partial^{\ell} a_{ij}\,\partial^{s}\lambda} = 0$ for any $\ell \geq 2$. In this case, equation (11) reduces to, after neglecting the terms containing more than (m+1) of (Δa_{ij}) and $(\Delta\lambda_k)$,

$$0 = \frac{\partial f}{\partial a_{ij}}\bigg|_{\lambda_k, a_{ij}} \Delta a_{ij} + \frac{1}{2}(\Delta\lambda_k)(\Delta a_{ij})\frac{\partial^2 f}{\partial\lambda\,\partial a_{ij}}\bigg|_{\lambda_k, a_{ij}} + \cdots + \frac{1}{(m-1)!}(\Delta a_{ij})(\Delta\lambda_k)^{m-1}\frac{\partial f^m(\lambda, a_{ij})}{\partial a_{ij}(\partial\lambda)^{m-1}}\bigg|_{\lambda_k, a_{ij}} + \frac{1}{m!}(\Delta\lambda_k)^m\frac{\partial^m f}{(\partial\lambda)^m}\bigg|_{\lambda_k, a_{ij}} \tag{13}$$

Hence if $f(\lambda, a_{ij})$ has pole $_k$ with multiplicity m, then $\Delta\lambda k/\Delta a_{ij}$ can be approximated by

$$\left|\frac{\Delta\lambda_k}{\Delta a_{ij}}\right| = \left|\frac{\dfrac{\partial f}{\partial a_{ij}} + (\Delta\lambda_k)\dfrac{\partial^2 f}{\partial a_{ij}\partial\lambda} + \ldots + \dfrac{1}{(m-1)!}(\Delta\lambda_k)^{m-1}\dfrac{\partial^m f}{\partial a_{ij}(\partial\lambda)^{m-1}}}{\dfrac{1}{m!}(\Delta\lambda_k)^{m-1}\dfrac{\partial^m f}{\partial\lambda^m}}\right|_{\lambda_k, a_{ij}} \tag{14}$$

which can be further approximated by, after neglecting the terms containing $(\Delta\lambda_k)$ in the numerator and assuming $\dfrac{\partial f}{\partial a_{ij}} \neq 0$,

$$\left|\frac{\Delta\lambda_k}{\Delta a_{ij}}\right| = \left|\frac{\dfrac{\partial f}{\partial a_{ij}}}{\dfrac{1}{m!}(\Delta\lambda_k)^{m-1}\dfrac{\partial^m f}{\partial\lambda^m}}\right|_{\lambda_k, a_{ij}} \tag{15}$$

Hence, if $\partial f/\partial a_{ij} \neq 0$, the sensitivity function for a polynomial containing a root λ_k with multiplicity m becomes

$$S^{\lambda_k}_{a_{ij}} = \left|\frac{a_{ij}}{\lambda_k}\frac{\Delta\lambda_k}{\Delta a_{ij}}\right| = \left|\frac{m! a_{ij}\dfrac{\partial f}{\partial a_{ij}}}{\lambda_k(\Delta\lambda_k)^{m-1}\dfrac{\partial^m f}{\partial\lambda^m}}\right|_{\lambda_k, a_{ij}} \tag{16}$$

This equation will be used only to establish a bound of $S_{a_{ij}}^{\lambda_k}$. If $S_{a_{ij}}^{\lambda_k}$ is to be computed explicitly, the m-th order polynomial of $\Delta\lambda$ in (14) has to be solved. Note that equation (16) can also be derived by using a different argument shown in [10]. We also note that in developing Eqs. (12),(16) it is assumed that Δa_{ij}, and consequently $\Delta\lambda_k$ are very small. In the following we shall further assume that $\Delta a_{ij} << a_{ij}$. Although we assume $\Delta a_{ij} << a_{ij}$, nothing can be said about $\Delta\lambda_k$ and λ_k; $\Delta\lambda_k$ can be larger or smaller than λ_k.

The sensitivity function (12) or (16) is due to the variation of the coefficient a_{ij}. Now in a simulation, all the coefficients which are different from 0 or 1 will cause the variation of λ_k. Hence the sensitivity of an eigenvalue λ_k in a simulation is defined as

$$S^{\lambda_k} = \sum_{\substack{a_{ij} \neq 0 \\ \downarrow \\ \neq 1}} \left| \frac{a_{ij}}{\lambda_k} \frac{\Delta\lambda_k}{\Delta a_{ij}} \right| \tag{17}$$

The elements 0 and 1 are excluded for the reasons that they are either open or direct connection in the simulation.

IV SENSITIVITY COMPARISONS OF THE JORDAN FORM AND THE NORMAL FORM SIMULATIONS

In this section we shall compare the eigenvalue sensitivities of the Jordan form and the normal form simulations of a differential equation. Let $S_N^{\lambda_k}$ and $S_J^{\lambda_k}$ be the sensitivity of the eigenvalue λ_k in the normal form simulation and the Jordan form simulation, respectively. We shall show that for each eigenvalue,

$$S_N^{\lambda_k} \geq S_J^{\lambda_k}$$

for any differential equation or a dynamical equation which may have repeated real and/or complex eigenvalues. We prove this first for the real eigenvalues and then the complex eigenvalues. Without loss of generality, it is assumed that the differential equation has a real eigenvalue λ_1 with multiplicity m and a pair of complex conjugate eigenvalues with multiplicity ℓ.

(a) Real eigenvalue

Because of the assumption, the Jordan form simulation is of the form shown in (6). It is easy to see that the sensitivity of real eigenvalue λ_1 is equal to 1, that is

$$S_J^{\lambda_1} = \sum_{\substack{a_{ij} \neq 0 \\ \downarrow \\ \neq 1}} \left| \frac{a_{ij}}{\lambda_1} \frac{\Delta\lambda_1}{\Delta a_{ij}} \right| = 1 \tag{19}$$

Note that this holds not only for the Jordan form simulation but also for any $\underline{A}$ matrix in the triangular form, because in this form, every element in

the diagonal is an eigenvalue, and the variation of each element in the diagonal alone has the effect on that eigenvalue.

In the normal form simulation, the dynamic equation is shown in Eq.(4). Its characteristic equation is

$$f(\lambda a_{ij}) = \lambda^n - a_1\lambda^{n-1} - a_2\lambda^{n-2} - \cdots - a_n$$
$$= (\lambda-\lambda_1)^m (\lambda^2 - d_1\lambda - d_2)^{\ell} = (\lambda-\lambda_1)^m \phi(\lambda\, b_i)$$

where $\phi(\lambda_1 b_i) = \lambda^{n-m} - b_1\lambda^{n-m-1} - \cdots - b_{n-m}$.

Since $\left.\frac{\partial f}{\partial a_i}\right|_{\lambda_1} = -\lambda_i^{n-i}$ and $\left.\frac{\partial^m f}{\partial \lambda^m}\right|_{\lambda_1} = m!\phi(\lambda_1, b_i)$, the sensitivity of λ_1 due to the parameter variations in the normal form can be computed as

$$S_N^{\lambda_1} = \sum_i \left| \frac{a_i \frac{\partial f}{\partial a_i}}{\frac{1}{m!}\lambda_1(\Delta\lambda_1)^{m-1}\frac{\partial^m f}{\partial \lambda^m}}\right|_{\lambda_1} = \frac{\left|-a_1\lambda_1^{n-1}\right|}{\left|\lambda_1(\Delta\lambda_1^{a_1})^{m-1}\phi(\lambda_1, b_i)\right|}$$
$$+ \frac{\left|-a_2\lambda_1^{n-2}\right|}{\left|\lambda_1(\Delta\lambda_1^{a_2})^{m-1}\phi(\lambda_1, b_i)\right|} + \cdots + \frac{\left|-a_n\right|}{\left|\lambda_1(\Delta\lambda_1^{a_n})^{m-1}\phi(\lambda_1, b_i)\right|} \qquad (20)$$

where $\Delta\lambda_1^{a_i}$ is the change of λ_1 due to the variation of a_i. In general, $\Delta\lambda_1^{a_i} \neq \Delta\lambda_1^{a_j}$ if $i \neq j$. In the following we shall show that $S_N^{\lambda_1} \geq S_J^{\lambda_1} = 1$. We consider only the case that $|\Delta\lambda_1^{a_i}| \leq \lambda_1$ for all i. Indeed if $|\Delta\lambda_1^{a_i}| > \lambda_1$ for some i, then $S_N^{\lambda_1} = \sum_i \left|\frac{\Delta\lambda_1/\lambda_1}{\Delta a_i/a_i}\right| > 1$ following the assumption $|\Delta a_i| \ll a_i$. Hence it is sufficient to consider the case $|\Delta\lambda_1^{a_i}| \leq \lambda_1$ for all i. In this case, we have

$$S_N^{\lambda_1} \geq \frac{\left|-a_1\lambda_1^{n-1}\right| + \left|-a_2\lambda_1^{n-2}\right| + \cdots + \left|-a_n\right|}{\left|\lambda_1(\lambda_1)^{m-1}\phi(\lambda_1, b_i)\right|}$$

and

$$S_N^{\lambda_1} - S_J^{\lambda_1} \geq \frac{\left|-a_1\lambda_1^{n-1}\right| + \left|-a_2\lambda_1^{n-2}\right| + \cdots + \left|-a_n\right|}{\left|\lambda_1^m\phi(\lambda_1, b_i)\right|} - 1$$

$$= \frac{(-a_1|\lambda_1|^{n-1} - a_2|\lambda_1|^{n-2} - \cdots - a_n) - (|\lambda_1^n - b_1\lambda_1^{n-1} - \cdots - b_{n-m}\lambda_1^m|)}{|\lambda_1^m(\lambda_1^{n-m} - b_1\lambda_1^{n-m-1} - \cdots - b_{n-m})|} \qquad (21)$$

Note that $(-a_i)$ and $(-b_i)$ are positive numbers following from the assumption that all the roots have negative real parts. Since

$$f(\lambda, a_i) = (\lambda-\lambda_1)^m(\lambda^{n-m} - b_1\lambda^{n-m-1} - \cdots - b_{n-m}\lambda^{1^m})$$

$$= (\lambda^m - b_1'\lambda^{m-1} - b_2'\lambda^{m-2} - \ldots - b_m')(\lambda^{n-m} - b_1\lambda^{n-m-1} - \ldots - b_{n-m})$$

$$= \lambda^n - a_1\lambda^{n-1} - a_2\lambda^{n-2} - \ldots - a_n$$

where $-b_1' = -m\lambda_1 > -\lambda_1 > 0$ and since $(-a_i)$, $(-b_i)$ and $(-b_i')$ are all positive, the following in equality between $(-a_i)$ and $(-b_i)$ can be obtained:

$$-a_1 = -b_1' - b_1 > -b_1 - \lambda_1$$

$$-a_2 = -b_2 + (-b_1')(-b_1) - b_2' > -b_2 + (-b_1')(-b_1) > -b_2 + (-\lambda_1)(-b_1)$$

and, in general,

$$-a_i > -b_i + (-\lambda_1)(-b_{i-1}).$$

for $i=3,4,\ldots,n-m$. By substituting $-a_i > -b_i + (-\lambda_1)(-b_i - 1)$ into (21) and observing $|\lambda_1^n - b_1\lambda_1^{n-1} - \ldots - b_{n-m}\lambda_1^m| \leq |\lambda_1^n| - |b_1||\lambda_1^{n-1}| - \ldots - |b_{n-m}||\lambda_1^m|$, (21) becomes

$$S_N^{\lambda_1} - S_J^{\lambda_1} \geq \frac{-(b_1+\lambda_1)|\lambda_1^{n-1}| - (b_2 - \lambda_1 b_1)|\lambda_1^{n-2}| - \ldots - (b_{n-m} - \lambda_1 b_{n-m-1})|\lambda_1^m|}{|\lambda_1^m(\lambda_1^{n-m} - b_1\lambda^{n-m-1} - \ldots - b_{n-m})|}$$

$$\frac{-|\lambda_1^n| + b_1|\lambda_1^{n-1}| + \ldots + b_{n-m}|\lambda_1^m|}{}$$

$$= \frac{-b_1|\lambda_1^{n-1}| - b_2|\lambda_1^{n-2}| - \ldots - b_{n-m-1}|\lambda_1^{m+1}}{|\lambda_1^m(\lambda_1^{n-m} - b_1\lambda^{n-m-1} - \ldots - b_{n-m})|} > 0$$

This proves that $S_N^{\lambda_1} \geq S_J^{\lambda_1}$.

(b) Complex conjugate eigenvalue: because of the limitation of the space, this part is omitted.

V. SENSITIVITY COMPARISON OF THE JORDAN FORM AND AN ARBITRARY FORM SIMULATIONS

As mentioned in Sec. II, in using the normal form and the (modified) Jordan form simulations, the number of components used in the simulation is much less than that used in an arbitary form simulation. Hence an arbitrary form simulation should not be used unless it is less sensitive to the parameter variations than the Jordan or normal form simulation. We show in this section that for the class of differential equations or dynamical equations with simple real eigenvalues, an arbitary form is <u>not</u> less sensitive to the parameter variations, hence it seems that there is no reason to use an arbitrary form simulation.

Consider a differential or dynamical equation with simple real eigenvalues. Let $S_J^{\lambda i}$ and $S_F^{\lambda i}$ be the sensitivity function of the Jordan form and an arbitrary form simulations, respectively. It is clear that $S_J^{\lambda i} = 1$ for each i. Let the matrix $\underline{A}$ in an arbitrary form simulation be of the form

$$\underset{\sim}{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},$$

and let its characteristic polynomial be

$$f(\lambda, a_{ij}) = \lambda^n - a_1\lambda^{n-1} - a_2\lambda^{n-2} - \ldots - a_n = \prod_{i=1}^{n}(\lambda - \lambda_i).$$

Then the coefficients a_i can be computed from the Faddeev algorithm[12] as, using the notation $\operatorname{tr} \underset{\sim}{A} = \sum_{i=1}^{n} a_{ii}$,

$$\underset{\sim}{A}_1 = \underset{\sim}{A}, \qquad a_1 = \operatorname{tr} \underset{\sim}{A}_1, \qquad \underset{\sim}{R}_1 = \underset{\sim}{A}_1 - a_1\underset{\sim}{I}$$

$$\underset{\sim}{A}_2 = \underset{\sim}{A}\,\underset{\sim}{R}_1, \qquad a_2 = \tfrac{1}{2}\operatorname{tr} \underset{\sim}{A}_2, \qquad \underset{\sim}{R}_2 = \underset{\sim}{A}_2 - a_2\underset{\sim}{I}$$

$$\vdots$$

$$\underset{\sim}{A}_n = \underset{\sim}{A}\,\underset{\sim}{R}_{n-1}, \qquad a_n = \frac{1}{n}\operatorname{tr} \underset{\sim}{A}_n, \qquad \underset{\sim}{R}_n = \underset{\sim}{A}_n - a_n\underset{\sim}{I} = 0.$$

From these relations, it can be shown that $\sum_{i,j} a_{ij} \frac{\partial a_k}{\partial a_{ij}} = k\, a_k$, $k=1,2,\ldots,n$.

Since the eigenvalues are assumed to be simple, $S_F^{\lambda_i}$ can be computed by using (12,17) as

$$S_F^{\lambda_i} = \sum_{i,j} \left| \frac{a_{ij} \frac{\partial f}{\partial a_{ij}}}{\lambda \frac{\partial f}{\partial \lambda}} \right|_{\lambda_i, a_{ij}} = \sum_{i,j} \left| \frac{a_{ij} \frac{\partial}{\partial a_{ij}}(\lambda^n - a_1\lambda^{n-1} - \ldots - a_n)}{\lambda \frac{\partial f}{\partial \lambda}} \right|_{\lambda_i, a_{ij}}$$

$$\geq \left| \frac{\sum_{i,j} a_{ij} \frac{\partial}{\partial a_{ij}}(-a_1\lambda^{n-1}) + \ldots + \sum_{i,j} a_{ij} \frac{\partial}{\partial a_{ij}}(-a_n)}{\lambda \frac{\partial f}{\partial \lambda}} \right|_{\lambda_i, a_{ij}}$$

$$= \frac{|-a_1\lambda_i^{n-1} - 2a_2\lambda_i^{n-2} - \ldots - na_n|}{\left|\lambda_i \prod_{\substack{j=1 \\ j\neq i}}^{n} (\lambda_i - \lambda_j)\right|}$$

Let $\prod_{\substack{j=1 \\ j\neq i}}^{n} (\lambda - \lambda_j) = \lambda^{n-1} - b_1\lambda^{n-2} - \ldots - b_{n-1}$. Then, from

$$f(\lambda, a_{ij}) = (\lambda - \lambda_i)(\lambda^{n-1} - b_1\lambda^{n-2} - \ldots - b_{n-1})$$

$$= \lambda^n - a_1\lambda^{n-1} - \ldots - a_n,$$

we obtain $a_1 = b_1 + \lambda_i$, $a_2 = b_2 - \lambda_i b_1, \ldots, a_{n-1} = b_{n-1} - \lambda_i b_{n-2}$ and $a_n = -\lambda_i b_{n-1}$. By substituting these into $S_F^{\lambda_i}$, we obtain

$$S_F^{\lambda_i} \geq \frac{|-(b_1+\lambda_i)\lambda_i^{n-1} - 2(b_2 - \lambda_i b_1)\lambda_i^{n-2} - \ldots - n(-\lambda_i b_{n-1})|}{|\lambda_i^n - b_1\lambda_i^{n-1} - \ldots - b_{n-1}\lambda_i|} = 1 = S_J^{\lambda_i}.$$

This proves that for the class of differential equations studied, an arbi-

trary form simulation is more sensitive to the parameter variation than the Jordan form simulation.

VI CONCLUSIONS

The problem of eigenvalue sensitivities of various simulations of differential or dynamical equations has been completely solved in this paper. The only condition imposed on the equations is that their eigenvalues have negative real parts, otherwise they may be simple or repeated, real or complex. It was prvoed that the Jordan form simulation is least sensitive to the parameter variations among all possible simulations. Hence if the sensitivity is an important factor in the simulation, the Jordan form dynamical equation should be used.

REFERENCES

[1] R. Vichneretsky,"Error Analysis in the Computer Simulation of Dynamic Systems: Variational Aspect of the Problem" IEEE Trans.on Electronic Computers, vol.EC-16,No.4,pp. 403-411.1967

[2] E.P.O'Grady,"Correlation Method for Computer Sensitivity Functions on a High-Speed Iterative Analog Computer" IEEE Trans. on Electronic Computers, vol. EC-16, No. 2, April, 1967

[3] Dr. J. Nelson, "A Fundamental Error Theory for Analog Computers" IEEE Trans. on Electronic Computer, Vol.EC-12, pp.541-550, 1963.

[4] V.A. Marsocci, "Dynamic Error in Electronic Analog Computers", IEEE Trans on Electronic Computers, Vol. EC-16, pp. 210-212, 1967.

[5] P.E. Mantrey, "Eigenvalue Sensitivity and State-Variable Selection", IEEE Trans. on Auto. Control, Vol. AC-13, pp. 263-269, June, 1968.

[6] C.K. Lee and C.T. Chen, "Sensitivity Comparisons of Analogue Computer" INT. J. Control, Vol. 7, 1968.

[7] R.A. Singer, "Selecting State Variables to Minimize Eigenvalue Sensitivity of Multivariable Systems". Automatica, vol.5, No. 1, pp. 85-94, 1969.

[8] D.C. Reddy, "Sensitivity of an Eigenvalue of a Multivariable Control System" Electr.Letters, vol. 2, p. 446, 1966.

[9] C.E. Maley, "The Effect of Parameters on the Roots of an Equation System", Computer J. , vol. 4, pp. 62-63, 1963.

[10] J.H. Wilkinson, Rounding Errors in Algebraic Processes, Prentice-Hall, Inc., Englewood Cliffs, N.J.

[11] B.S. Morgan, Jr., "Sensitivity Analysis and Synthesis of Multivariable Systems", IEEE Trans. on Auto. Control, Vol. AC-11, pp.506,July, 1966.

[12] D.K. Faddeev and V.N. Faddeeva, Computational Methods of Linear Algebra, W.H. Freeman and Company, San Franciso, 1963.

SINGULAR LINE COMPENSATORS

B. L. SHAPIRO
Major, U.S. Marine Corps
Naval Postgraduate School
Monterey, California

G. J. Thaler
Naval Postgraduate School
Monterey, California

INTRODUCTION

It is well known that the dynamics of a linear system are modeled by a characteristic polynomial, and that the roots of this polynomial are functions of the parameters of the system. Using Siljak's[1,2] methods the root-parameter relationships can be portrayed on the α-β parameter plane. The usual procedure maps chosen lines from the s-plane onto the parameter plane, and for most cases the mapping is unique, i.e., each point in the s-plane maps into one point (or at worst a finite number of points) in the parameter plane. This means, of course, that adjustment of parameters to obtain a desired pair of complex roots requires a unique pair (or pairs) of parameter values.

The relationship between roots and parameters is expressed as a pair of simultaneous algebraic equations. These are linear equations if the pertinent parameters appear in the polynomial coefficients in linear algebraic form, and are nonlinear equations if the parameter relationships are nonlinear. In either case, for certain values of s it is possible for the simultaneous algebraic equations to be linearly dependent. This condition was first explored in detail by Bowie[3]. When linear dependence exists an infinite number of solutions are possible for the simultaneous albebraic equations, so the value of s which causes this condition maps into a line on the parameter plane. Such lines have been called "Singular Lines." Their physical interpretation is that roots of the characteristic polynomial may be obtained at the designated s value with an infinite number of pairs of parameter values, i.e., any pair which are coordinates of a point on the singular line.

It is apparent that such singular line conditions, when obtainable for desired values of s, can be used to alleviate adjustment problems and some tolerance problems. It has also been shown (Desrosiers, Thaler[4]) that singular lines can be used in adaptive systems.

One remaining problem is that of designing the system to have a singular line at a specified value of s. This paper presents one method for designing compensators to produce desired singular lines.

THEORETICAL BACKGROUND

Given the characteristic polynomial

$$F(s) = \sum_{k=0}^{n} a_K s_k = 0 \tag{1}$$

where $a_k = b_k\alpha + c_k\beta + d_k$, and $s = -\zeta\omega_n \mp j\omega_n\sqrt{1-\zeta^2}$, one requires the real and imaginary parts of (1) to go to zero independently and obtains

$$B_1\alpha + C_1\beta = -D_1 \; ; \quad B_2\alpha + C_2\beta = -D_2 \tag{2}$$

where B_1, C_1, D_1, B_2, C_2, D_2, are defined in Appendix I.

The coefficient and augmented matrices for (2) are

$$A = \begin{bmatrix} B_1 & C_1 \\ B_2 & C_2 \end{bmatrix} \qquad AH = \begin{bmatrix} B_1 & C_1 & -D_1 \\ B_2 & C_2 & -D_2 \end{bmatrix} \tag{3}$$

from which, for those cases in which the common rank of A and AH is two:

$$\alpha = \frac{C_1D_2 - C_2D_1}{B_1C_2 - B_2C_1} \qquad \beta = \frac{B_2D_1 - B_1D_2}{B_1C_2 - B_2C_1} \tag{4}$$

For those cases in which Eq. (2) are linearly dependent, matrix A is singular rank one and det A = 0. To obtain a solution one must first find a ζ, ω_n pair such that matrices A and AH have common rank one, then one of Eq. (2) provides the solution, which is seen to be a linear relationship between α and β and thus plots as a straight line on the α and β parameter plane. All points on the line are maps of the s-plane point (ζ_s, ω_{ns}), thus the straight line is a constant ζ, constant ω_n line.

These results, by themselves, have not been convenient for design purposes. If the parameter plane concept is extended to a three parameter space, the parametric simultaneous equations become

$$\text{(a)} \quad B_1\alpha + C_1\beta + D_1\gamma + F_1 = 0; \quad \text{(b)} \quad B_2\alpha + C_2\beta + D_2\gamma + F_2 = 0 \tag{5}$$

which when solved give the line of intersection of two planes in 3-parameter space. A singular line in the parameter plane may be obtained by rotating the line of Eq. (5) into any plane parallel to the α-β plane.

DESIGN OF COMPENSATION TO PRODUCE SINGULAR LINES

An expedient way to rotate the singular line is to note that γ must be some function of α and β. Then Eq. (5(b)) may be multiplied by D_1/D_2 providing

$$\frac{D_1B_2}{D_2}\alpha + \frac{D_1C_2}{D_2}\beta + D_1\gamma + \frac{D_1F_2}{D_2} = 0 \tag{6}$$

Adding (5(a)) and (6), and solving for γ gives

$$\gamma = -\frac{1}{2}\left[\left(\frac{B_1}{D_1}+\frac{B_2}{D_2}\right)\alpha+\left(\frac{C_1}{D_1}+\frac{C_2}{D_2}\right)\beta+\frac{F_1}{D_1}+\frac{F_2}{D_2}\right] \tag{7}$$

Substituting into (5(a)) and (6),

$$\begin{aligned}(D_2B_1-D_1B_2)\alpha+(D_2C_1-D_1C_2)\beta+D_2F_1-D_1F_2=0\\(D_1B_2-D_2B_1)\alpha+(D_1C_2-D_2C_1)\beta+D_1F_2-D_2F_1=0\end{aligned} \tag{8}$$

Equations (8) differ only by a factor of -1, so they are obviously linearly dependent. Selection of the desired value of s for the singular line, and substitution in the appropriate equations evaluates B,C,D and F which are then substituted into Eq. (7) to find γ.

To design compensators which cause the system to have appropriate singular lines, a compatible compensation scheme is chosen, and a parameter (frequently a gain) of that compensator is designated $\gamma(\alpha,\beta)$. The characteristic polynomial is then derived and the procedures previously indicated are applied.

EXAMPLES

The type 0 system of Fig. 1 is to be compensated by adjusting the feedback gain, γ, and it is desired to select γ such that there are roots at

$$s = -\zeta\omega_n \mp j\omega_n\sqrt{1-\zeta^2}\quad,\quad \zeta = 0.5,\quad \omega_n = 1.0.$$

The characteristic equation is

$$s^3+(\alpha+\beta+8)s^2+(8\alpha+4\beta+\gamma+16)s+15\alpha+4\gamma+4=0 \tag{9}$$

With the indicated coefficients, B,C,D,F are calculated:

$$B_1=-14\qquad C_1=1\qquad D_1=-4\qquad F_1=3$$

$$B_2=-7\qquad C_2=-3\qquad D_2=-1\qquad F_2=-8$$

Substitution in Eq. (7) gives

$$\gamma = -\frac{21}{4}\alpha-\frac{11}{8}\beta-\frac{29}{8} \tag{10}$$

From which the singular line is

$$14\alpha+13\beta+35=0 \tag{11}$$

The characteristic equation may now be revised to

$$s^3+(\alpha+\beta+8)s^2+\left(\frac{11\alpha}{4}+\frac{21\beta}{8}+\frac{94}{8}\right)s-6\alpha-\frac{11\beta}{2}-\frac{21}{2}=0 \tag{12}$$

which has all roots in the LHP if $\alpha \leqq 56$ and β is determined from Eq. (11).

Fig. 2 shows a third order Type 1 plant compensated with velocity feedback and a cascaded filter. The characteristic equation becomes

$$s^4 + (\alpha+1+k_2k_1)\, s^3 + [(1+k_2k_1)\alpha + \beta + k_2k_1]s^2 +$$

$$[k_2k_1(\alpha+\beta)+4\gamma]s + 4k_1\gamma = 0 \qquad (13)$$

Again choose $\zeta = 0.5$ and $\omega_n = 1.0$ for the singular line, then

$B_1 = k_2k_1 \qquad C_1 = 1 \qquad D_1 = -4k_2 \qquad F_1 = -1$

$B_2 = 1 \qquad C_2 = -k_2k_1+1 \qquad D_2 = -4 \qquad F_2 = k_2k_1-1$

The singular line becomes

$$(-4k_2k_1+4k_1)\alpha + (-4-4k_2k_1^2+4k_1)\beta + 4+4k_2k_1^2 - 4k_1 = 0 \qquad (14)$$

If we choose $k_2 = 10$ to provide a spanning of one decade between pole and zero, Eq. (14) becomes

$$-36k_1\alpha + (-40k_1^2 + 4k_1 - 4)\beta + 40k_1^2 - 4k_1 + 4 = 0 \qquad (15)$$

Then the compensation function is

$$\gamma = \frac{1}{8k_1}[11k_1\alpha + (1+k_1-10k_1^2)\beta - 1 - k_1 + 10k_1^2] \qquad (16)$$

which gives a characteristic equation

$$s^4 + (\alpha+1+10k_1)s^3 + [(1+10k_1)\alpha + \beta + 10k_1]s^2$$

$$+ [(10k_1+\frac{11}{2})\alpha + (5k_1 + \frac{1}{2k_1} + \frac{1}{2})\beta - \frac{1}{2k_1} + 5k_1 - \frac{1}{2}]s \qquad (17)$$

$$+ \frac{11k_1}{2}\alpha + (\frac{1}{2} + \frac{k_1}{2} - 5k_1^2)\beta - \frac{1}{2} + 5k_1^2 - \frac{k_1}{2} = 0$$

Choosing $k_1 = 2$ to obtain dominance for the singular roots, the singular line becomes

$$6\alpha + 13\beta - 13 = 0 \qquad (18)$$

and the characteristic equation becomes

$$s^4 + (\alpha+21)s^3 + (21\alpha+\beta+20)s^2$$

$$+ (\frac{51\alpha}{2} + \frac{43\beta}{4} + \frac{37}{4})s + 11\alpha + \frac{37\beta}{4} + \frac{37}{2} = 0 \qquad (19)$$

while the required compensator becomes

$$G_c = \frac{(\frac{8}{11}\alpha - \frac{37}{16}\beta + \frac{37}{16})(s+12)}{s+20} \qquad (20)$$

A parameter plane plot showing the singular line is given in Fig. 3.

IDENTIFICATION AND SELF ADAPTATION

When used in a self-adaptive loop [4] an identification scheme that identifies parameter changes seems desirable. Once the change of the parameter has been determined in magnitude and sense, this information is used to change the second (or adjustable) parameter to return to the singular line and maintain the chosen complex roots at the specified location.

This adjustment is an open-loop, calibrated adjustment, since the equation of the singular line is known.

Indirect identification and closed loop adjustment should be possible. If a parameter variation causes a shift in root location, identification methods may be applied to the root variation rather than the parameter variation. Assuming a singular line design a second parameter which is adjustable is available, and is known to be capable of accomplishing the desired adaptation. It should then be possible to make this parameter adjustment with a closed loop scheme that uses root identification to detect the error, i.e., the deviation of the root from its desired location. This possibility has not yet been explored.

DISCUSSION AND CONCLUSIONS

A simple, direct method for the design of singular line compensators has been presented. This makes possible designs which alleviate some tolerance problems, and offers possibilities for design of self adaptive systems.

APPENDIX I

Theory of the Parameter Plane Method

Let the characteristic equation be

$$F(s) = a_n s^n + a_{n-1} s^{n-1} + \cdots + a_2 s^2 + a_1 s + a_o = 0 \tag{1}$$

in which the coefficients are in general, non-linear functions of the two system parameters α and β and s is the complex variable. Let s be expressed as

$$s = -\omega_n \zeta + j\omega_n \sqrt{1-\zeta^2} \tag{2}$$

where ω_n is the undamped natural frequency and ζ is the relative damping coefficient and substitute

$$\theta = \cos^{-1}(-\zeta) \tag{3}$$

There results

$$s = \omega_n(\cos\theta + j \sin\theta) = \omega_n e^{j\theta} \tag{4}$$

Then s^k may be written

$$s^k = \omega_n{}^k e^{jk\theta} = \omega_n{}^k \ (\cos k\theta + j \sin k\theta) \tag{5}$$

Now substituting

$$T_k(-\zeta) = \cos k\theta = \cos(k\cos^{-1}-\zeta)$$

$$U_k(-\zeta) = \frac{\sin k\theta}{\sin\theta} = \frac{\sin(k\cos^{-1}-\zeta)}{\sin(\cos^{-1}-\zeta)} \tag{6}$$

there results

$$s^k = \omega_n{}^k[T_k(-\zeta) + j\sqrt{1-\zeta^2}\, U_k(-\zeta)] \tag{7}$$

It may be shown that

$$T_k(-\zeta) = (-1)^k T_k(\zeta) \tag{8}$$

$$U_k(-\zeta) = (-1)^{k+1} U_k(\zeta)$$

where the functions $T_k(\zeta)$ and $U_k(\zeta)$ are Chebyshev functions of the first and second kinds, respectively. Eq. (7) therefore becomes

$$s^k = \omega_n{}^k[(-1)^k T_k(\zeta) + \sqrt{1-\zeta^2}(-1)^{k+1} U_k(\zeta)\ ] \tag{9}$$

Substituting in Eq. (1) and equating the summations of the real and imaginary parts independently to zero results in the two equations,

$$\begin{aligned} &\sum_{k=0}^{n} a_k \omega_n{}^k (-1)^k T_k(\zeta) = 0 \\ &\sum_{k=0}^{n} a_k \omega_n{}^k (-1)^{k+1} U_k(\zeta) = 0 \end{aligned} \tag{10}$$

The functions $T_k(\zeta)$ and $U_k(\zeta)$ may be obtained by applying the recurrence formulae

$$\begin{aligned} &T_{k+1}(\zeta) - 2\zeta T_k(\zeta) + T_{k-1}(\zeta) = 0 \\ &U_{k+1}(\zeta) - 2\zeta U_k(\zeta) + U_{k-1}(\zeta) = 0 \end{aligned} \tag{11}$$

with

$$T_o(\zeta) = 1,\ T_1(\zeta) = \zeta, U_o(\zeta) = 0,\ U_1(\zeta) = 1,\ U_{-1}(\zeta) = -1$$

Also the functions $T_k(\zeta)$ may be expressed in terms of the functions $U_k(\zeta)$ by means of the relationship

$$T_k(\zeta) = \zeta U_k(\zeta) - U_{k-1}(\zeta) \tag{12}$$

and upon substitution in Eq. (10) the following simultaneous equations are obtained.

$$\begin{aligned} &\sum_{k=0}^{n} (-1)^k a_k \omega_n{}^k U_{k-1}(\zeta) = 0 \\ &\sum_{k=0}^{n} (-1)^k a_k \omega_n{}^k U_k(\zeta) = 0 \end{aligned} \tag{13}$$

Eq. (13) can be considered as two equations in the two unknowns α and β, which may be rearranged and solved for α and β as

$$\alpha = \alpha(\omega_n, \zeta) \qquad \beta = \beta(\omega_n, \zeta) \tag{14}$$

In rearranging the equations the specific format necessary depends upon the

coefficient-parameter relationships. In general, the algebraic form of the coefficients will correspond to one of the following cases.

$$\text{I.} \quad a_k = b_k\alpha + c_k\beta + d_k \tag{15}$$

$$\text{II.} \quad a_k = b_k\alpha + c_k\beta + h_k\alpha\beta + d_k \tag{16}$$

$$\text{III.} \quad a_k = b_{k2}\alpha^2 + b_{k1}\alpha + h_k\alpha\beta + c_{k1}\beta + d_k \tag{17}$$

$$\begin{aligned}\text{IV.} \quad a_k = {} & b_{kn}\alpha^n + b_{k(n-1)}\alpha^{n-1} + \cdots \\ & \cdots + h_{k(n-1)}\alpha^{n-1}\beta + \cdots \\ & \cdots + c_{kn}\beta^n + c_{k(n-1)}\beta^{n-1} + \cdots \\ & + d_k\end{aligned} \tag{18}$$

In the following derivations only the first case is treated. Solutions also have been obtained for the other cases, and the derivations for these are included in Ref. 3.

Case I

$$a_k = b_k\alpha + c_k\beta + d_k$$

For this case Eq. (13) rearrange into the form

$$\begin{aligned}\alpha B_1(\zeta,\omega_n) + \beta C_1(\zeta,\omega_n) + D_1(\zeta,\omega_n) = 0 \\ \alpha B_2(\zeta,\omega_n) + \beta C_2(\zeta,\omega_n) + D_2(\zeta,\omega_n) = 0\end{aligned} \tag{19}$$

where

$$B_1 = \sum_{k=0}^{n} (-1)^k b_k \omega_n^k U_{k-1}(\zeta)$$

$$B_2 = \sum_{k=0}^{n} (-1)^k b_k \omega_n^k U_k(\zeta)$$

$$C_1 = \sum_{k=0}^{n} (-1)^k c_k \omega_n^k U_{k-1}(\zeta)$$

$$C_2 = \sum_{k=0}^{n} (-1)^k c_k \omega_n^k U_k(\zeta)$$

$$D_1 = \sum_{k=0}^{n} (-1)^k d_k \omega_n^k U_{k-1}(\zeta)$$

$$D_2 = \sum_{k=0}^{n} (-1)^k d_k \omega_n^k U_k(\zeta) \tag{20}$$

Using Cramer's rule solutions for α and β are

$$\alpha = \frac{C_1D_2 - C_2D_1}{B_1C_2 - B_2C_1} \qquad \beta = \frac{B_2D_1 - D_2B_1}{B_1C_2 - B_2C_1} \tag{21}$$

REFERENCES

[1] Siljak, D. D., "Analysis and Synthesis of Feedback Control Systems in the Parameter Plane," Part I, II, III, IEEE Transactions on Applications and Industry, v. 83, Nov 1964.

[2] Siljak, D. D., "Generalization of the Parameter Plane Method," IEEE Transactions on Automatic Control, V. AC-11, Jan 1966.

[3] Bowie, D. J. and G. J. Thaler, "Singular Lines in the Parameter Plane," Proc. of the Second Annual Princeton Conference on Information, 1968.

[4] Desrosiers, R. A. and G. J. Thaler, "Self-Adaptation Using Singular Lines in the Parameter Plane," Second IFAC Symposium on Sensitivity and Self-Adaptivity, Dubrovnik, Yugoslavia, 1968.

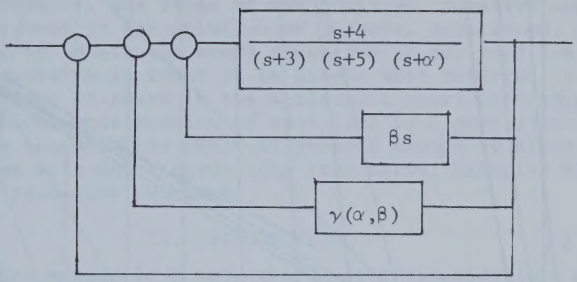


Figure 1.

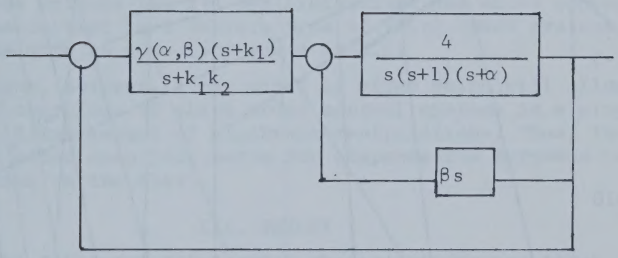


Figure 2.

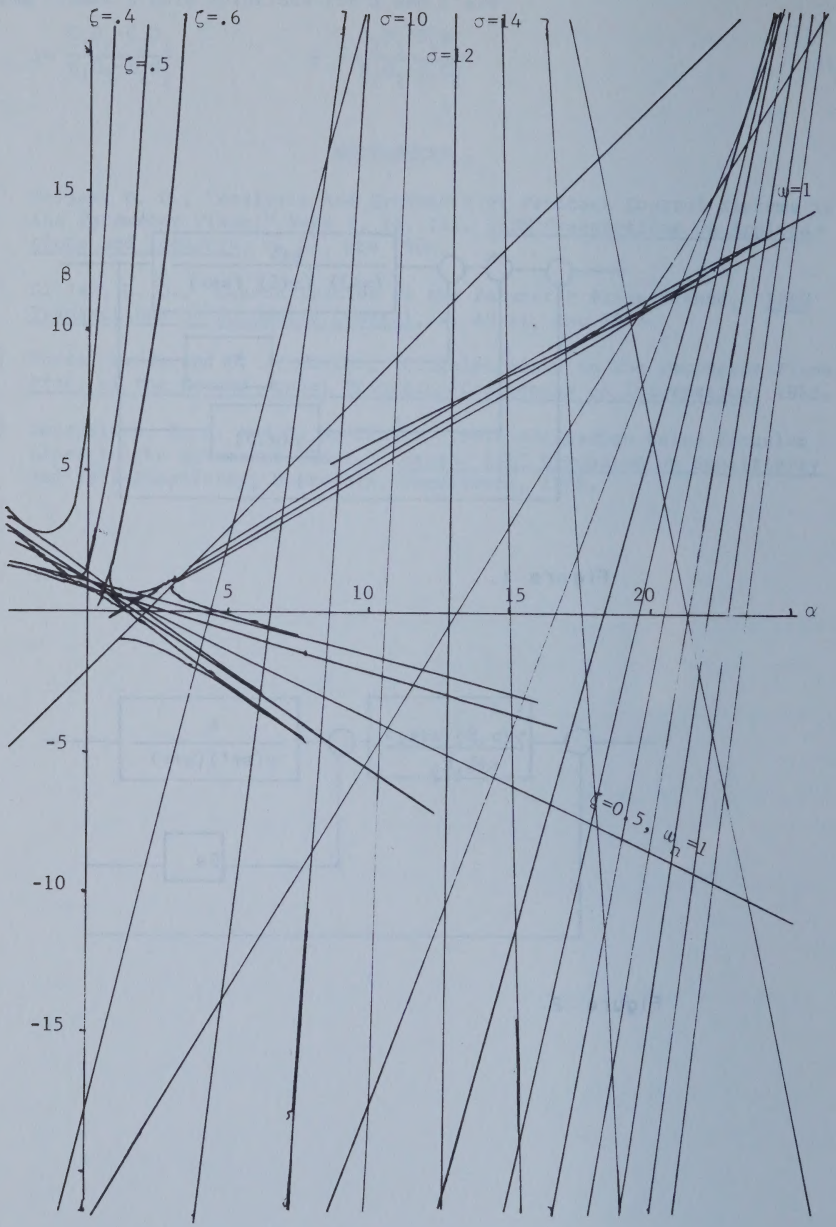


Fig. 3. PARAMETER PLANE CURVES SHOWING SINGULAR LINE

A NEW CHART FOR ROOT LOCUS CONSTRUCTION OF GENERALIZED THIRD ORDER CONTROL SYSTEMS

M. Z. Mohammed PMIEEE & Dr. Engineer A. Y. Belal

Both are with the Electrical Engineering Department,
University of Mosul, Iraq.

I. ABSTRACT

Third order control system, with and without open loop zeros are considered. Specifically, the ideas of the positive, negative and zero sequences, usually used in the solution of 3-phase, unbalanced, power system are adopted to obtain the exact equations of the root locus of the system. This equation is found to be simple when referred to some fixed axis at 120° out of phase in the s-plane. A chart is then given, allowing the graphical construction of root loci in a very straight forward and simple manner. This chart is general enough to allow studying control systems with and without open loop zeros. Examples are given to illustrate the procedure proposed.

II. INTRODUCTION

The third order system is a very useful approximation to a great many practical control systems. The root locus diagram gives a very useful picture of how the closed loop poles location vary as open-loop-parameters such as gain K are varied. Apart from the normal classical methods, there has appeared in the literature, geometrical methods[1], and methods using a generalized third order loci[2, 3]. Concerning the special problems dealt with here in this paper, the authors have devised a graphical procedure based on the so called (p-q) and (β^2-q) charts[4]. Each of the above methods has its own limitations and short comings, and some of them specialized in a certain type of third order systems, or consider systems with no open loop zeros at all.

In this paper, however, a new chart is given which will allow the construction of root loci of third order control systems in a simple manner with a minimum amount of algebraic manipulations. Thus, the effect of extra added open loop zeros for compensation purposes can be easily manipulated on the chart.

III. THEORY

Consider the following third order characteristic equation:

$$s^3 + As^2 + Bs + C = 0 \quad \ldots\ldots (1)$$

where A, B and C are constants.

Let:

$$s = x - A/3 \quad \ldots\ldots (2)$$

using (2) into (1) gives:

$$x^3 + px + q = 0 \quad \ldots\ldots (3)$$

where:

$$p = B - A^2/3$$
$$q = A^3/27 - BA/3 + C \quad \ldots\ldots (4)$$

The location of two of the 3 roots of equation (1) is shown reffered to

the s and x planes in figure (1). Thus for the sake of definition, if:

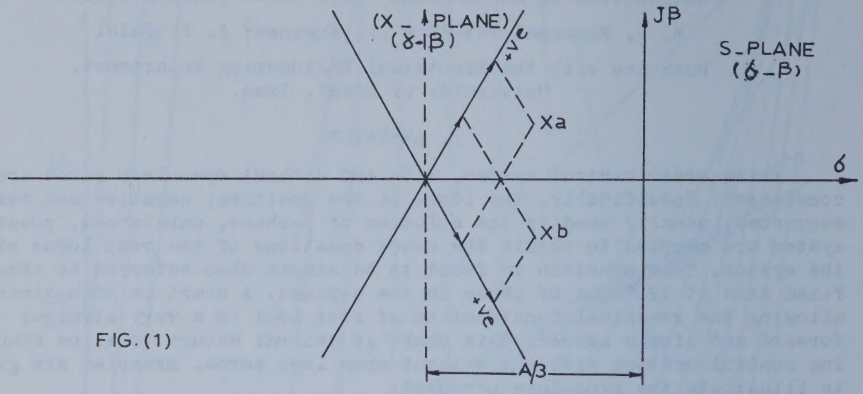


FIG. (1)

$$s = \delta + j\beta \quad \ldots\ldots (5)$$

then

$$x = \gamma + j\beta \quad \ldots\ldots (6)$$

where

$$\gamma = A/3 + \delta \quad \ldots\ldots (7)$$

In the x-plane, the 3 roots x_a, x_b, x_c represent an unbalanced 3 phase system of vectors with sequence a, b, c, where:

$$\begin{aligned} x_a &= \gamma + j\beta \\ x_b &= \gamma - j\beta \\ x_c &= -2\gamma + j0 \end{aligned} \quad \ldots\ldots (8)$$

The above 3 unbalanced vectors can be decomposed of 3 sets of balanced system of vectors (5), namely:

Positive sequence (a, b, c): (x_{a1}, s_{b1}, x_{c1})

Negative sequence (a, c, b): (x_{a2}, x_{b2}, x_{c2})

Zero sequence: (x_{ao}, x_{bo}, x_{co})

where the following relationships are true:

$$\begin{aligned} x_{ao} &= 1/3\,(x_a + x_b + x_c) \\ x_{a1} &= 1/3\,(x_a + ax_b + a^2x_c) \\ x_{a2} &= 1/3\,(x_a + a^2x_b + ax_c) \end{aligned} \quad \ldots\ldots (9)$$

The symbol a defines an operator as follows:

$$\begin{aligned} a &= 1\,\angle 120^\circ = -\tfrac{1}{2} + j\sqrt{3}/2 \\ a^2 &= 1\,\angle 240^\circ = -\tfrac{1}{2} - j\sqrt{3}/2 = \bar{a} \end{aligned} \quad \ldots\ldots (10)$$

with the bar denoting the complex conjugate.

Substituting from (8) and (10) into (9) gives:

$$x_{ao} = 0$$

$$x_{a1} = 1/3 \left[- 3 \bar{a} \gamma + j\beta (1 - a) \right] \quad \text{.....(11)}$$

$$x_{a2} = 1/3 \left[- 3 a \gamma + j\beta (\overline{1 - a}) \right]$$

Substituting for a, $\bar{a}$, (1 - a) and $(\overline{1 - a})$, equation (11) gives:

$$x_{a1} = (\frac{\gamma}{2} + \frac{\beta}{2\sqrt{3}}) + j\sqrt{3} (\frac{\gamma}{2} + \frac{\beta}{2\sqrt{3}})$$

$$x_{a2} = (\frac{\gamma}{2} - \frac{\beta}{2\sqrt{3}}) - j\sqrt{3} (\frac{\gamma}{2} - \frac{\beta}{2\sqrt{3}}) \quad \text{..... (12)}$$

Equations (12) show that, for all values of and , the positive vector x_{a1} is always inclined at 60° with the -axis in the x-plane. Also the negative sequence vector x_{a2} is always inclined at -60° with the -axis in that plane. Multiplying the left hand side of equation (12), gives:

$$x_{a1} x_{a2} = \gamma^2 - \frac{\beta^2}{3} \quad \text{..... (13)}$$

It is shown in the appendix that:

$$\gamma^2 - \frac{2}{3} = - \frac{P}{3} \quad \text{..... (14)}$$

Substituting from equation (14) into (13), gives:

$$x_{a1} x_{a2} = - P/3 \quad \text{..... (15)}$$

For a given value of p defined by (4), equation (15), defines the exact root locus upon which the pole x_a lie referred to the fixed axis of the positive and negative sequence vectors in the x-plane.

Again, equation (12) gives:

$$x_{a1}^3 + x_{a2}^3 = (2\gamma)^3 + (2\gamma)p \quad \text{..... (16)}$$

It is also shown in the appendix that:

$$(2\gamma)^3 + 2\gamma p = q \quad \text{..... (17)}$$

where q is defined by equation (4).

Substituting (17) into (16) gives:

$$x_{a1}^3 + x_{a2}^3 = q \quad \text{..... (18)}$$

using (18) and (15), the location of x_a, hence the three roots of equation (1) are defined.

IV. ROOT LOCUS CHART CONSTRUCTION

Case (a) Open loop transfer function with no zeros: Consider the following system:

$$G(s) = \frac{K}{s^3 + As^2 + Bs + C} \quad \text{..... (19)}$$

where K is the open loop gain $0 \leq K \leq \infty$

The closed loop characteristic equation is as follows:

$$s^3 + As^2 + Bs + (C + K) = 0 \quad \ldots\ldots (20)$$

using transformation (2), equation (20) gives:

$$x^3 + p_1 x + q_1 = 0 \quad \ldots\ldots (21)$$

where:

$$\begin{aligned} p_1 &= p \\ q_1 &= q + K \end{aligned} \quad \ldots\ldots (22)$$

p and q are as has been defined by equation (4). Using the results of the theory of previous paragraph, gives:

$$x_{a1}\, x_{a2} = - P_1/3 = - p/3 \quad \ldots\ldots (23)$$

$$x_{a1}^3 + x_{a2}^3 = q_1 = q + K \quad \ldots\ldots (24)$$

Thus equation (23) gives the equation of the root locus referred to the positive and negative sequence coordinates in the x-plane, and equation (24) can be used to calibrate the root locus in terms of the system gain K.

From the above, it is concluded that, knowing the A, B, and C parameters of 3rd order control systems of the form given by equation (19) defines p from (4) and a corresponding root locus in the x-plane (or the shifted s-plane). The chart of figure (2) shows a family of such loci for different values of p positive or negative. The calibration family of curves corresponding to equation (24) is drawn on the same chart for different values of K. Notice that for K = 0, the points on the root locus corresponds to open loop poles of the original open loop system assumed not easily available from equation (19).

It should be noticed that points on the real axis could be calibrated in terms of K by using equation (21) and (22) combined which give:

$$K = - q - px - x^3 \quad \ldots\ldots (21)$$

Case (b), Open loop transfer functions with one zero: In this case, open loop systems of the following form are considered:

$$G(s) = \frac{K(s + z)}{s^3 + As^2 + Bs + C} \quad \ldots\ldots (25)$$

where (- z) is an extra open loop zero.

The closed loop characteristic equation is given by:

$$s^3 + As^2 + (B + K)s + (C + Kz_1) = 0 \quad \ldots\ldots (26)$$

Transformation (2) gives:

$$x^3 + p_2 s + q_2 = 0 \quad \ldots\ldots (27)$$

where

$$p_2 = p + K$$
$$q_2 = q + K\,(z_1 - A/3) \qquad \ldots\ldots (28)$$

The root locus and its calibration in terms of the system gain K, can now be defined by:

$$x_{a1}\, x_{a2} = -\, p_2/3 = -\,(p + K)/3 \qquad \ldots\ldots (29)$$

$$x_{a1}^3 + x_{a2}^3 = q_2 = q + K\,(z_1 - A/3) \qquad \ldots\ldots (30)$$

For each value of K, p_2 and q_2 are evaluated defining one locus from (29) and another locus from (30) on the chart of figure (2). Therefore, the characteristic roots of (26) are directly plotted on the chart.

Also, calibration of the real axis x could be made by making use of equation (27) and (28) which give:

$$K = -\,\frac{q - px - x^3}{x + z_1 - A/3} \qquad \ldots\ldots (27)$$

Case (c): Open loop transfer functions with two zeros: Consider an open loop transfer function with the following rather general form:

$$G(s) = \frac{K\,(Ds^2 + Es + F}{s^3 + As^2 + Bs + C} \qquad \ldots\ldots (31)$$

where D, E, and F define the two zeros of the open loop system. It is to be noted that 3 parameters are used while only two are sufficient so as to include the previous two special cases in the present case.

Hence when D = 0 case (b) will result, and when both D and E = 0 case (a) results. The closed loop characteristic equation takes the following form:

$$s^3 + (A + KD)\, s^2 + (B + KE)\, s + (C + KF) = 0 \qquad \ldots\ldots (32)$$

Making the transformation:

$$x = s + \frac{A + KD}{3} \text{ gives} \qquad \ldots\ldots (33)$$

$$x^3 + p_3 x + q_3 = 0 \qquad \ldots\ldots (34)$$

where:

$$p_3 = p + a_1 K + a_2 K^2$$
$$q_3 = q + b_1 K + b_2 K^2 + b_3 K^3 \qquad \ldots\ldots (35)$$

In equation (35), the a, s and b, s are defined as:

$$a_1 = E - 2\,A\,D/3$$
$$a_2 = -\,D^2/3$$
$$b_1 = F - (AE + BD)/3 + (2/9)\, A^2 D \qquad \ldots\ldots (36)$$
$$b_2 = (2/9)\, D^2 A - (ED)/3$$
$$b_3 = 2\, D^3/27$$

As before, by a similar arguments, the root locus with its points calibrated in terms of K, is defined in the x-plane by:

$$x_{a_1} x_{a_2} = - p_3 = - (p + a_1 K + a_2 K^2)/3 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (37)$$

$$x_{a_1}^3 + x_{a_2}^3 = q_2 = q + b_1 K + b_2 K^2 + b_3 K^3 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (38)$$

Following the same previous procedure, the characteristic roots of (32) can thus be defined in the x-plane. The situation now is somewhat different because, transformation (32) has to be used for each point corresponding to each value of K if the root locus in the s-plane is to be obtained.

V. EXAMPLES

Example one: Consider the multiloop system of figure (3).

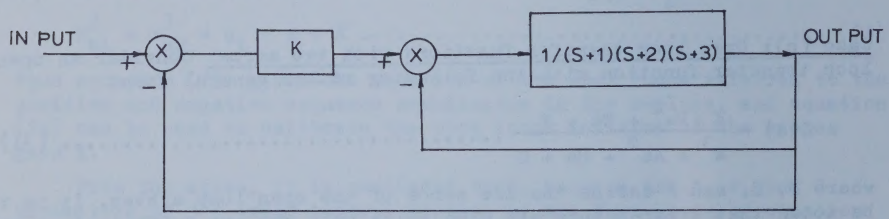


FIG. (3)

The open loop transfer function of this system is given by:

$$G = \frac{K}{s^3 + 6 s^2 + 11 s + 7} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (39)$$

where

$A = 6$, $B = 11$, and $C = 7$.

Using transformation (2) gives $x = 2 + s$

equation (4) and (22) give:

$$p_1 = - 1 \quad \& \quad q_1 = 1 + K$$

Therefore, the root locus lies on the constant p-curve of $p_1 = - 1$ starting at the point where $q_1 = 1$.

The values of gain K along the locus are fixed by the points of intersection with the constant q curves as shown in figure (2).

Example two: As another example, a second transfer function of the form

$$G = \frac{K}{s^3 + s - 1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (40)$$

Following the same procedure as before with the present values of $p_1 = 1$ and $q_1 = K - 1$, the root locus of this transfer function can be easily plotted and is shown on the chart of figure (2). Values of K are also shown.

Example three: A third example is now given to cover another shape of third order root locus diagram. Therefore it is considered the following transfer function:

$$G = \frac{K}{s^3 - 2s - 2} \qquad \ldots\ldots (41)$$

In this case:

$$p_1 = -2$$

$$q_1 = -2 + K.$$

Using these values for different values of K, the root locus is plotted as it is shown in fig. (2).

From this figure it is shown that the roots are complex for the region of K starting at the poles corresponding to K = 0 up to the break in point where K = 0.8, and for other values of K higher than the break away point with K = 3.2.

That part on the real axis is calibrated using equation (21).

Example four: Consider the case of one zero:

$$G = \frac{K(s + 1)}{s^3 - s - 2} \qquad \ldots\ldots (42)$$

equation (28) gives.

$$p_2 = K - 1 \quad \text{and} \quad q_2 = K - 2 \qquad \ldots\ldots (43)$$

Substituting for some values for K in equations (43) the following table is obtained:

K	0	1	2	3	4	5
p	-1	0	1	2	3	4
q	-2	-1	0	1	2	3

Locating the above values of p and q on the chart, the root locus is constructed in a straight forward manner and is shown in figure (2).

VI. CONCLUSIONS AND REMARKS

A method has been presented to construct root loci of general third order control systems with and without open loop zeros, not necessarily given in the usual factored form. The method is very simple and straight forward, with a minimum amount of algebraic manipulations.

As a result of analytical investigation using the ideas of symmetrical components, a chart has been given enabling one to consider and design such systems based on a graphical procedure.

Following the procedure of this paper, future developments concerning higher order control systems may be achieved.

VII. REFERENCES

1. N. G. Headows: "Generalized root locii for 3rd order control systems"
   Control, pp. 198 - 201, April 1965.

2. T. Konwerski: "A synthesis of 3rd order type 1 system"
   Control pp. 316 - 321, April 1968.

3. D. J. Bullen: "The determination of parameters of a 3rd order system to give a spedified response"
   International Journal of Electrical Engineering Education, Vol. 5, No. 1, pp. 67 - 70, June 1967.

4. M. Z. Mohammed and A. Y. Belal: "A method of synthesis of third order control systems",
   sent for publication in the IEEE Transaction on Automatic Control.

5. Stevenson: "Elements of power system analysis"
   McGraw-Hill.

6. S. E. Williamson: "Design data to assist the plotting of root loci"
   Control pp. 556 - 558, June 1968.

VIII. APPENDIX

A proof of the formulae $\frac{\beta^2}{3} = \gamma^2 + p/3$ and $q = (2\gamma)^3 + (2\gamma)p$

Let:

$$x = \gamma + j\beta \tag{1-A}$$

Substituting (1-A) into (3) gives:

$$\gamma^3 - 3j\gamma^2\beta - 3\gamma\beta^2 - \beta^3 j + p\gamma + jp\beta + q = 0 \tag{2-A}$$

Equating real and imaginary parts to zero gives:

$$\gamma^3 - 3\gamma\beta^2 + p\gamma + q = 0 \tag{3-A}$$

and

$$3\gamma^2 - \beta^2 + p = 0$$

or

$$\frac{\beta^2}{3} = \gamma^2 + p/3 \tag{4-A}$$

Substituting from (4-A) into (3-A) gives:

$$(-2\gamma)^3 + (-2\gamma)\,p + q = 0$$

or

$$q = (2\gamma)^3 + (2\gamma)\,p \tag{5-A}.$$

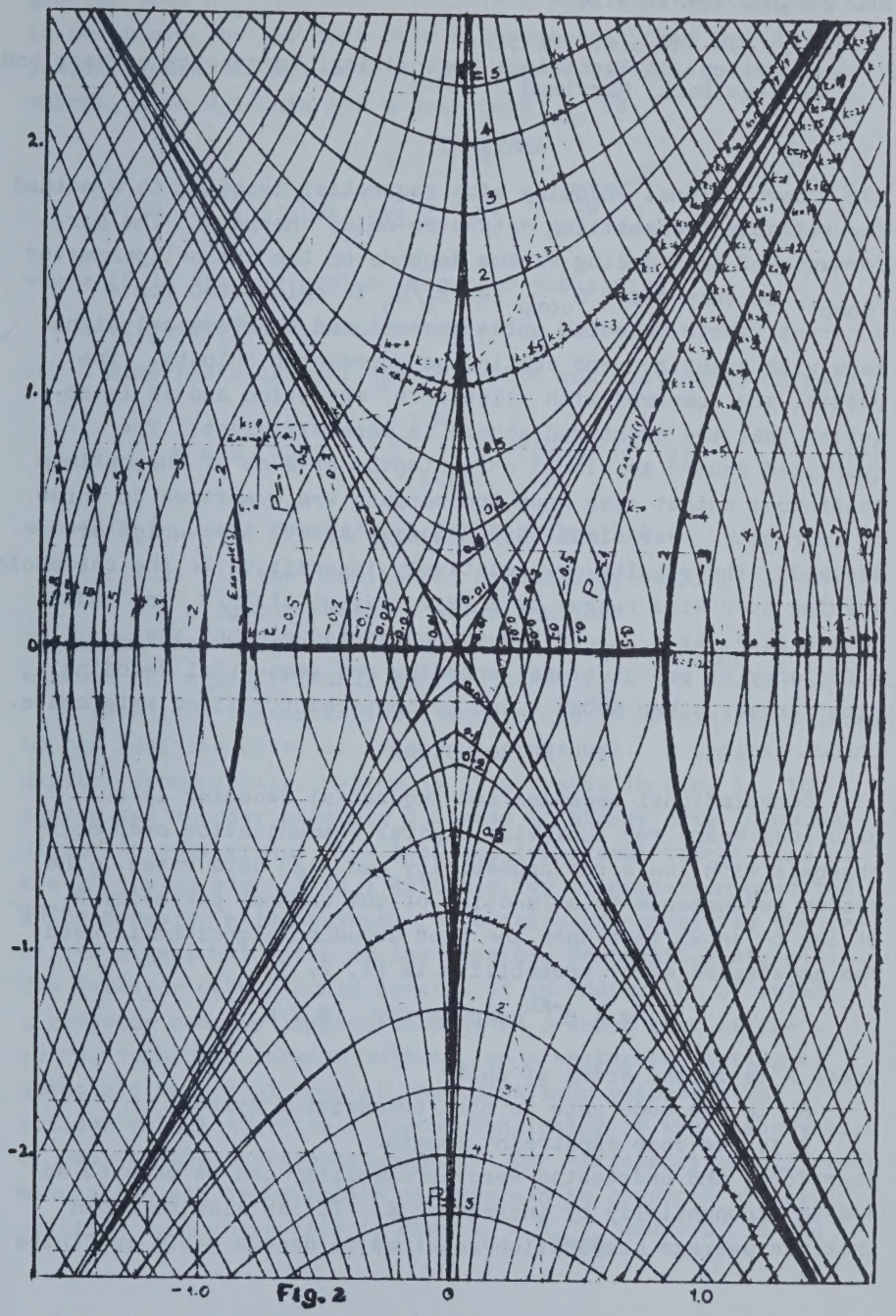


Fig. 2

PERFORMANCE OF SEQUENTIAL DECODING WITH BIORTHOGONAL MODULATION AND Q-LEVEL QUANTIZATION

M.C. KIM AND S.A. TRETTER
Department of Electrical Engineering, University of Maryland, College Park, Maryland

ABSTRACT

Convolutional encoding with sequential decoding is a method for reliably transmitting data over noisy channels. The performance of this coding scheme depends on the channel parameter R_{comp}. In this paper R_{comp} vs. E_b/N_o is calculated for a coherent, white Gaussian noise channel and biorthogonal signal sets with quantized and continuous correlator outputs. The results are compared with "list of L" detection and it is observed that the quantized scheme is more efficient. The quantized scheme is within a few tenths of a db of the optimum continuous output case when the outputs are quantized to three or four bits. For simplicity, equally spaced thresholds are assumed. The results are relatively insensitive to the threshold spacing over wide ranges of E_b/N_o. Curves of E_b/N_o for code rate R equal to R_{comp} vs.1/R are presented so that the efficiency of convolutional encoding and sequential decoding with various other MODEM's can be compared for fixed rate codes.

I. INTRODUCTION

Convolutional encoding and sequential decoding is a practical means for achieving digital communication over noisy channels with low error probability and high data rates. The system performance is a function of the channel parameter R_{comp}. It can be shown that when the Fano decoding algorithm is used, the undetected error probability is [1, 2, 3]

$$p_e \leq A\, 2^{-KR_{comp}/R_N} \text{ for } R_N \leq R_{comp}$$

where A is in the order of unity

R_N is the code rate in bits/transmitted symbol

K is the code constraint length.

Therefore, the undetected error probability can be made to decrease exponentially by increasing K. It can also be shown that the average computation required to decode a bit is finite

for code rates less than R_{comp} and unbounded for code rates greater than R_{comp}. Thus, to achieve real time decoding it is necessary for the code rate R_N to be less than the channel computational cut-off rate R_{comp}. For a discrete memoryless channel with M inputs and Q outputs R_{comp} is

$$R_{comp} = \max_{\{P_k\}} \left\{ -\log_2 \sum_{i=1}^{Q} \left[\sum_{k=1}^{M} P_k \sqrt{q_{ik}} \right]^2 \right\} \qquad (1)$$

where $\{P_k\}$ are the apriori input probabilities

$\{q_{ik}\}$ are the channel transition probabilities.

For a memoryless channel with M inputs and continuous outputs Y

$$R_{comp} = \max \left\{ -\log_2 \int_{-\infty}^{\infty} \left[\sum_{k=1}^{M} P_k \sqrt{f(Y/k)} \right]^2 dY \right\} \qquad (2)$$

where $\{P_k\}$ are the apriori input probabilities

$f(Y/k)$ is the conditional probability density for the output vector Y given input k.

It can be shown [3] that for symmetric channels R_{comp} is a maximum for equally likely inputs, i.e., $P_k = 1/M$ for $k=1,\ldots,M$.

Satellite communication channels can be accurately modelled by the additive white Gaussian noise channel. A typical coded digital communication system is illustrated in Fig. 1. The source and encoder generate a stream of binary data. The modulator accepts groups of k bits and transforms them to one of $M = 2^k$ analog signals $\{s_k\}$ suitable for transmission over the white Gaussian noise channel. The detector correlates the received signal with each of the possible transmitted signals. The detector output could consist of an indication of which correlator had the largest output, an ordered list of the L correlators with largest outputs, or a vector consisting of quantized values of each of the correlator outputs. The optimum output would be a vector consisting of the exact values of each of the correlator outputs. The efficiency of different modulation and demodulation schemes can be compared by calculating E_b/N_o necessary to make $R_N = R_{comp}$ where E_b is the

energy transmitted per information bit and N_o is the one sided noise spectral density.

In this paper R_{comp} is calculated for biorthogonal signal sets with quantized and continuous correlator outputs. The results are compared with "list of L" detection [4]. A set of M biorthogonal signals consists of a set of M/2 orthogonal signals and their negatives. Let $\{a_i(t)\}_{i=1}^{M/2}$ be a set of orthogonal signals, each with energy E_N and duration T. The biorthogonal signal set $\{s_i(t)\}$ can be indexed as follows:

$$s_i(t) = a_i(t) \qquad \text{for} \qquad 1 \le i \le M/2$$

$$s_i(t) = -a_{i-M/2}(t) \quad \text{for} \quad M/2 + 1 \le i \le M$$

If the received signal is $y(t) = s_k(t) + n(t)$ where $n(t)$ is white Gaussian noise with spectral density $N_o/2$, the normalized output of correlator i is

$$y_i = \frac{1}{\sqrt{\frac{N_o E_N}{2}}} \int_0^T y(t)\, s_i(t)\, dt.$$

Since $s_i(t) = -s_{i+M/2}(t)$ for $1 \le i \le M/2$

it is clear that only M/2 correlators are necessary. The probability density for y_i given $s_k(t)$ transmitted is

$$f(y_i/s_k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-b)^2}{2}} \qquad i = k$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i+b)^2}{2}} \qquad i = k \pm M/2 \text{ and } 1 \le i \le M$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{y_i^2}{2}} \qquad \text{otherwise}$$

where $b = \sqrt{\frac{2E_N}{N_o}}$

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_{M/2} \end{bmatrix}$ be the correlator output vector. Since the noise is white and Gaussian the elements of Y are statistically independent and

$$f(Y/s_k) = \prod_{i=1}^{M/2} f(y_i/s_k)$$

II. R_{COMP} WITH CONTINUOUS CORRELATOR OUTPUTS

From Equation 2 and the symmetry of biorthogonal signal sets it follows that for exact correlator outputs R_{comp} is given by

$$R_{comp} = -\log_2 \left\{ \frac{1}{M^2} \int_{-\infty}^{\infty} \left[\sum_{k=1}^{M/2} \sqrt{f(Y/s_k)} + \sum_{k=\frac{M}{2}+1}^{M} \sqrt{f(Y/s_k)} \right]^2 dY \right.$$

This expression reduces to

$$R_{comp} = \log_2 \frac{M}{1+\exp(-E_N/N_o)+(M-2)\exp(-E_N/2N_o)} \qquad (3)$$

Jordan [5] has found that with orthogonal signal sets, coherent detection, and continuous outputs

$$R_{comp} = \log_2 \frac{M}{1+(M-1)\exp(-E_N/2N_o)} \qquad (4)$$

Graphs of Equations 3 and 4 show that the biorthogonal signal sets are slightly more efficient in terms of E_N/N_o. In addition, they require half the bandwidth of the orthogonal sets.

III. R_{COMP} WITH Q-LEVEL QUANTIZATION

In practice the correlator outputs must be quantized due to finite accuracy constraints. It will be assumed that each correlator output is quantized into one of Q states. The

detector output is then the quantized correlator output vector

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{M/2} \end{bmatrix} \quad \text{where } \hat{y}_i = k \text{ for } y_i \in \Delta_k$$

for $i = 1, \ldots, M/2$ and $k = 1, \ldots, Q$

with $\Delta_k \cap \Delta_i = \Phi$ for $k \neq i$

and $\bigcup_{k=1}^{Q} \Delta_k = \{-\infty, \infty\}$

Thus the channel output $\hat{Y}$ can take on $Q^{M/2}$ different states. According to Equation 1 and the symmetry of the signal set

$$R_{comp} = -\log_2 \left\{ \frac{1}{M^2} \sum_{i_1=1}^{Q} \cdots \sum_{i_{M/2}=1}^{Q} \left[\sum_{k=1}^{M} \sqrt{P\{\hat{y}_1 = i_1, \ldots, \hat{y}_{M/2} = i_{M/2}/s_k\}} \right]^2 \right\}.$$

Since the correlator outputs are independent

$$P\{\hat{y}_1 = i_1, \ldots, \hat{y}_{M/2} = i_{M/2}/s_k\} = \prod_{j=1}^{M/2} P\{\hat{y}_j = i_j/s_k\}.$$

and

$$P\{\hat{y}_j = i_j/s_k\} = \int_{\Delta_{i_j}} f(y_j/s_k)\, dy_j$$

where $f(y_j/s_k)$ is the conditional correlator output probability density given in Section I.

Let $P\{\hat{y}_j = i_j/s_k\} = \begin{cases} P_{s+n}(i_j) & j = k \\ P_{-s+n}(i_j) & j = k - M/2 \\ P_n(i_j) & j \neq k \text{ or } k - M/2 \end{cases}$

Then the expression for R_{comp} can be simplified to

$$R_{comp} = \log_2 \frac{M}{1 + \sum_{k=1}^{Q} \sqrt{P_{s+n}(k) P_{-s+n}(k)}}$$

$$+\frac{1}{2}\left(\frac{M}{2} - 1\right) \left[\sum_{k=1}^{Q} \left(\sqrt{P_{s+n}(k) P_n(k)} + \sqrt{P_{-s+n}(k) P_n(k)} \right) \right]^2 \quad (5)$$

For simplicity equally spaced thresholds were chosen. The thresholds were arranged symmetrically about 0 as shown in Figure 2. It was found that the optimum threshold spacing, i.e., the spacing that maximized R_{comp} varied with E_N/N_o. However, the results varied little over wide ranges of E_N/N_o. In order to generate M-ary signals with R_N bits per signal the binary encoder must have a code rate of $R = R_N/\log_2 M$. The calculations are summarized in Figures 3 - 7. E_b/N_o vs. $1/R$ for $R_N = R_{comp}$ is plotted for various M and Q. $E_b = E_N/R_N$ is the energy used per information bit. Curves for "list of L" detection are also presented for comparison [4].

IV. CONCLUSIONS

From the curves it can be seen that with as little as three or four bits quantization (Q = 8 or 16) the performance of the quantized scheme is within several tenths of a db of the optimum continuous output case. List of L detection can not approach the continuous output case even for a full list due to the loss of information. According to Teoste [6] the results are similar when orthogonal signal sets are used. Teoste concludes that when a general purpose digital computer is used for decoding, quantization is more efficient computationally than list decoding. The results show that for coherent channels biorthogonal signal sets are more efficient than orthogonal signal sets in terms of E_N/N_o. For large M the two are nearly equivalent. However, the biorthogonal set requires half the bandwidth of the orthogonal set.

REFERENCES

1. J.E. Savage, "Progress in Sequential Decoding," Advances in Communication Systems, vol. 3, Academic Press Inc., New York, 1968.

2. J.M. Wozencroft and I.M. Jacobs, Principles of Communication Engineering, John Wiley & Sons, New York, 1965, Ch. 6.

3. R.G. Gallager, Information Theory and Reliable Communication, John Wiley & Sons, New York, 1968.

4. I.M. Jacobs, "Sequential Decoding with Biorthogonal Alphabet and List Decoding," JPL Space Programs Summary No. 37-33, vol. IV, April 1 - May 31, 1965, pp. 299-301.

5. K.L. Jordan, "The Performance of Sequential Decoding in Conjunction with Efficient Modulation," IEEE Transactions on Communication Technology, vol. com.-14, No. 3, June 1966, pp. 283-297.

6. R. Teoste, "M-ary Channel Quantization for Sequential Decoding," Lincoln Laboratory, M.I.T., JA2817, June 28, 1966.

ACKNOWLEDGEMENTS

This research was partially supported by a Faculty Research Award from the General Research Board of the University of Maryland. It was stimulated by Dr. Tretter's association with the IBM Center for Exploratory Studies, Rockville, Md., during the summer of 1968.

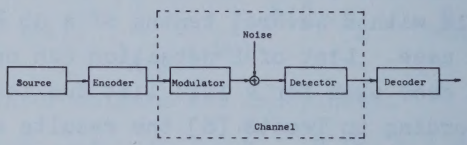


Fig. 1 A Typical Digital Communication System

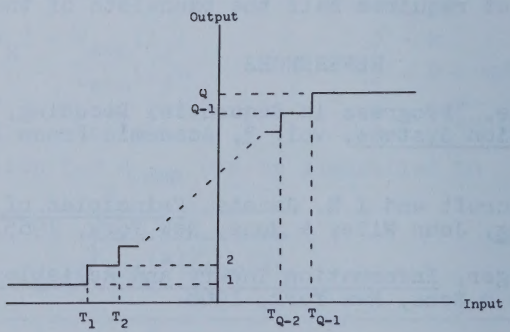


Fig. 2 Quantizer Input-Output Relationship

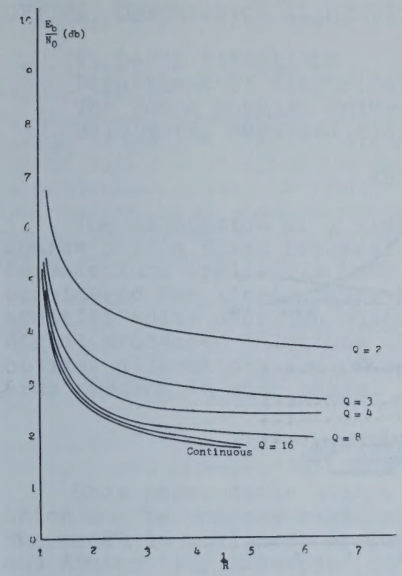


Fig. 3 E_b/N_0 vs. 1/R for $R_N = R_{comp}$ and M=2

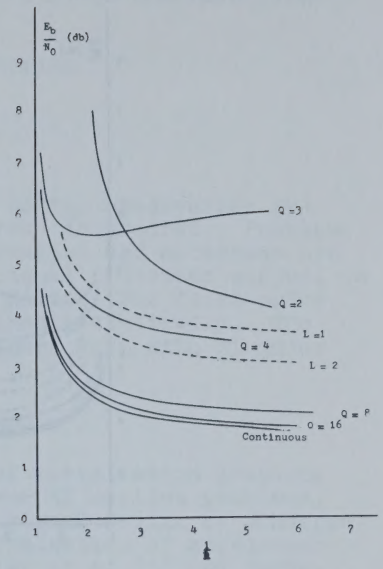


Fig. 4 E_b/N_0 vs. 1/R for $R_N = R_{comp}$ and M=4

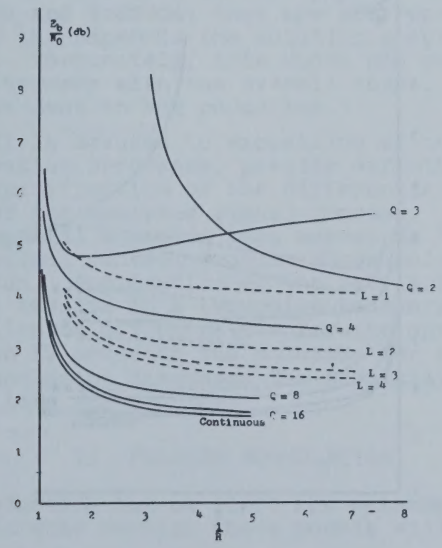


Fig. 5 E_b/N_0 vs. 1/R for $R_N = R_{comp}$ and M=8

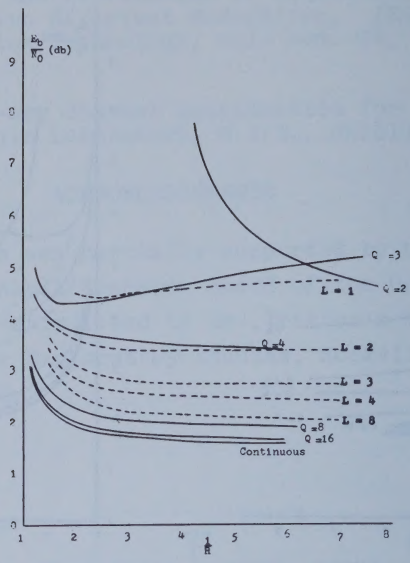


Fig. 6 E_b/N_o vs. 1/R for $R_N = R_{comp}$ and M=16

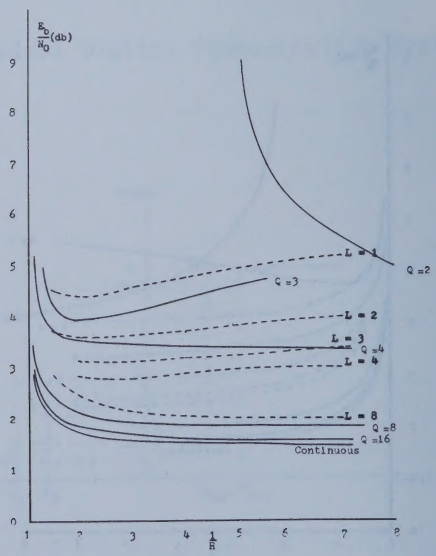


Fig. 7 E_b/N_o vs. 1/R for $R_N = R_{comp}$ and M=32

OPTIMAL OBSERVATION ALLOCATION IN ESTIMATION AND DETECTION

V. DAVID VANDELINDE
Department of Electrical Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

ABSTRACT

The allocation of a limited or costly observation resource over a fixed processing interval is studied. Problem formulations applicable both to estimation and detection are considered for linear Gauss-Markov signal processes subject to additive white Gaussian observation noise. For first-order signal processes solutions are obtained analytically. The optimal allocations are of the bang-bang form with singular arcs present in some cases.

I. INTRODUCTION

This paper deals with a class of optimization problems which may be characterized as resource allocation problems. The model is motivated by papers on signal design by Schweppe and Athans [1], Schweppe [2] and optimization of measurement subsystems in optimal control by Meier et al.[3] and Vande-Linde[4]. The basic problem is one of allocating a costly or limited observing capacity during a fixed processing interval in a manner to "best" estimate the state of the signal process. The above cited papers show that many signal design and observation allocation problems can be reduced to the model that is discussed in this paper. As such reductions can be involved and tedious, they are not reproduced. Rather, the concern of the paper is the solution properties of the resulting model. Fortunately, this model has an interpretation of its own in harmony with the overall theme, hence little intuitive feel is lost in the reduction.

Section II is devoted to exposition of the model of signal and observation processes, precise definition of "best" estimate and specification of the differential equation for the variance of the observed signal process. This matrix Ricatti differential equation then serves as the state equation in the optimal control problem whose solution is the desired allocation. Application of the matrix Maximum Principle in Section III results in a two-point boundary value problem and the so-called on-off principle for the optimal allocation [1,4]. Section IV presents the solution for the scalar problem and the conclusion discusses the possibility of further investigation into the problem.

II. PROBLEM FORMULATION

The Introduction has outlined the mathematical models to be studied. In this section these models will be made precise.

The State Equations

Consider a dynamical system whose state may be expressed as an n-vector $x(t)$ described by the vector differential equation

$$\dot{x}(t) = Ax(t)+Bv(t) \tag{2.1}$$

where

A is a nxn matrix
B is a nxs matrix
v(t) is a s-vector.

Let $x(0)$ be the state of the system at time 0. Assume that $x(0)$ is distributed according to a Gaussian law with mean value zero and non-singular covariance matrix $P(0)$. Also, assume that $v(t)$ is a vector white noise process with zero mean and with covariance matrix

$$\mathrm{Cov}[v(t);v(\tau)] = V\delta(t-\tau) = \mathcal{E}[v(t)v'(\tau)] \tag{2.2}$$

with V a positive-definite symmetric sxs matrix. Further, let the v process be uncorrelated with $x(0)$.

The Observation Equations

Let the observable variables be a p-vector $y(t)$, given as linear combinations of the state variables plus noise

$$y(t) = (u(t))^{\frac{1}{2}} Hx(t)+w(t) \tag{2.3}$$

where

H is a pxn matrix of full rank
w(t) is a p-vector
u(t) is a scalar.

Assume that $w(t)$ is a vector white noise process with zero mean and covariance matrix

$$\mathrm{cov}[w(t);w(\tau)] = W\delta(t-\tau) \tag{2.4}$$

where W is a pxp positive definite symmetric matrix. Moreover, assume that the w process is independent of the state. The observation control or allocation variable $u(t)$ is assumed bounded to the unit interval, i.e.

$$u(t) \in [0,1], \text{ for } t \in [0,T]. \tag{2.5}$$

The square root appears in (2.3) for notational convenience.

The Performance Criteria

Two performance criteria are considered. The first weights the estimation error only at the terminal time, T, and the second weights it throughout the observation interval. Define the nxn matrix $P(t)$ by

$$P(t) = \mathcal{E}\Big[(x(t)-\mathcal{E}[x(t)|y(\tau)\ 0\le\tau\le t])(x(t)-\mathcal{E}[x(t)|y(\tau)\ 0\le\tau\le t])' \Big| y(\tau)\ 0\le\tau\le t\Big] \quad (2.6)$$

That is, P(t) is the conditional covariance matrix of the state. The performance criterion for case 1 is given by

$$J_1(u) = \varphi(P(T))+\alpha\int_0^T u(t)dt \quad (2.7)$$

where $\varphi:R^{n^2}\to R$ is a differentiable convex function and $\alpha>0$ is a real scalar. For case 2

$$J_2(u) = \int_0^T tr(FP(t))+\alpha u(t)dt \quad (2.8)$$

where $tr(\cdot)$ is the trace function and F is a $n\times n$ positive semi-definite matrix. In either case an optimal allocation, u^*, is a time function which minimizes the performance criterion. The variable α may be thought of either as the cost of observing or as a constant Lagrange multiplier adjoining an integral constraint on the observation effort to the cost of incorrect estimation.

The Covariance Equation

The model described above has a Gaussian conditional probability distribution and fits into the framework of the Kalman-Bucy theory of filtering [5]. Thus the conditional covariance obeys a matrix Ricatti equation.

$$\dot{P}(t) = AP(t)+P(t)A'+BVB'-u(t)P(t)H'W^{-1}HP(t). \quad (2.9)$$

This equation establishes the link between the allocation function u and the covariance P. Henceforth it will serve as the state equation in the optimization problem. For notational symplicity define the matrices

$$C = BVB' \ , \ D = H^1W^{-1}H. \quad (2.10)$$

III. APPLICATION OF THE MAXIMUM PRINCIPLE

One of the most useful tools of modern systems theory is the Maximum Principle of Pontryagin [6]. This technique allows the replacement of an optimal control problem with a two-point boundary value differential system. In [1,4] a matrix version of the Maximum Principle is given and this version is applicable to the current problem. Before stating the resulting boundary value problem, it is convenient to define the following quantities. The adjoint differential equation is defined as

$$\dot{Z}(t) = \begin{cases} (DPu-A')Z+Z(PDu-A) & \text{for } J_1 \\ (DPu-A')Z+Z(PDu-A)+F & \text{for } J_2 \end{cases} \tag{3.1}$$

and the switching function

$$\psi(P,Z) = \mathrm{tr}[ZPDP]+\alpha \tag{3.2}$$

Next, applying the Maximum Principle obtains: If u^* is the optimal function, it is necessary that

$$u = \begin{cases} 1 & \text{if } \psi(P,Z) < 0 \\ 0 & \text{if } \psi(P,Z) > 0 \\ ? & \text{if } \psi(P,Z) = 0 \end{cases} \tag{3.3}$$

where Z and P are defined by (3.1) and (2.9) subject to the given initial condition P(0) and the terminal condition

$$Z(T) = \begin{cases} -\nabla_P \phi(P(T)) & \text{for } J_1 \\ 0 & \text{for } J_2 \end{cases} \tag{3.4}$$

The most important equation in this section is (3.3), for it contains both a result and a question. If ψ is not equal to zero except in passing, then the optimal modulation is of the bang-bang or on-off category [1]. However, there is the possibility of ψ equalling zero over an interval of positive length. In such a case (3.3) gives no information to determine u^*. Situations in which this occur are called singular [7] and other techniques must be used to find u^* during a singular period. In the next section such a situation arises and is acted upon.

IV. THE ONE-DIMENSIONAL PROBLEM

In this section the problem is solved for the scalar version, that is for n=s=p=1. Both case 1 and case 2 are studied with case 1 being the simpler of the two. All of the previous variable names are retained with lowercase letters replacing uppercase letters to emphasize that the variables are scalars.

Solution to Case 1

The problem is to find u^* such that

$$J_1 = \phi(p(T))+\alpha\int_0^T u(t)dt \tag{4.1}$$

is minimized; subject to

$$\dot{p} = 2ap+c-up^2d \tag{4.2}$$

with p(0) given. This is reduced by Section III to finding u^* to satisfy

$$u^* = \begin{cases} 1 & \psi(p,z) < 0 \\ 0 & \psi(p,z) > 0 \\ ? & \psi(p,z) = 0 \end{cases}$$

subject to

$$\begin{aligned} \psi(p,z) &= zp^2d+\alpha \\ \dot{p} &= 2ap+c-up^2d \\ \dot{z} &= 2(pdu-a)z \end{aligned} \tag{4.4}$$

with $p(0)$ given and $z(T) = \nabla_p\varphi(p(T))$. In this section $\varphi(p(T))$ is taken as the identity function when switching times are computed. The problem solution is clearest if it is solved separately for $a \geq 0$ and $a<0$.

$\underline{a \geq 0}$. In the pz-plane consider the curve $\psi(p,z) = 0$. Observe that trajectories are restricted to the fourth quadrant and $u = 0$ above the switching curve and $u = 1$ below the curve. To determine the conditions for singular trajectories take the inner product of the gradient of $\psi(p,z)$ and the time derivatives of the trajectories. This indicates the direction in which trajectories cross $\psi(p,z) = 0$.

$$\langle(\dot{p},\dot{z})', \nabla\psi(p,z)\rangle = 2pdz(ap+c) \tag{4.5}$$

This term is strictly less than zero unless p or z equal zero, as $c > 0$ and $d > 0$. Since $\nabla\psi(p,z)$ points into the $u = 0$ region of the space, all switches in u^* will be to $u^* = 1$. As the inner product is never zero and does not change sign, no singular condition can arise and at most one switch can occur.

Therefore

$$u^*(t) = \begin{cases} 0 & 0 \leq t \leq t_1 \\ 1 & t_1 < t \leq T. \end{cases} \tag{4.6}$$

Solving the boundary value problem (4.4) using this allocation function gives t_1 as the solution of

$$\alpha = [dp^2(0)e^{4at_1}-c]/[\cosh f(T-t_1)+\sinh f(T-t_1)$$

$$([dp(0)e^{2at_1}-a]/f)]^2 \tag{4.7}$$

where $f = -(a^2+dc)^{\frac{1}{2}}$. If $c = 0$ this simplifies to

$$\alpha = a^2/d[ae^{-a(T-t_1)}/(p(o)d)+\sinh a(T-t_1)]^2. \tag{4.8}$$

The latter equation is not as complicated as it first appears since it is quadratic in e^{at_1}.

$a<0$. Following the same procedure as before, again gives (4.5), only this time the term can be made to equal zero by setting $p = -c/a$. The geometric situation is shown in figure 1.

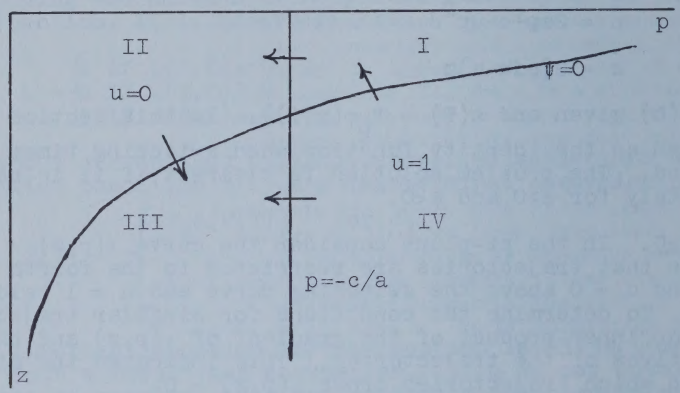


Figure 1. The pz plane for case 1 with $a<0$.

To the left of the line $p = -c/a$ trajectories cross into the region numbered III and to the right into I. Also, at $p = -c/a$, $\dot{p} < 0$ and $\dot{z} > 0$ so that the line is crossed from region I to region II and region IV to region III. The singular point at the intersection of $p=-c/a$ and the switching curve cannot generate a singular solution because $\dot{p} < 0$ at this point for any value of u. Noting the possible sequences of switching the optimal allocation u^* is given by

$$u^*(t) = \begin{cases} 1 & 0 \le t \le t_1 \\ 0 & t_1 \le t \le t_2 \\ 1 & t_2 < t \le T. \end{cases} \qquad (4.9)$$

This does not imply that all three intervals must be non-empty for a given problem. Determining the values of t_1 and t_2 is best carried out by a simple computer program. If $c = 0$, then $t_2 = T$ and t_1 is the solution of

$$\alpha = 4a^2e^{2a(T+t_1)}/d[2a/(p(o)r)-1+e^{2at_1}]^2. \qquad (4.10)$$

Qualitatively, these solutions are very appealing. Stability ($a < 0$) tends the solution to observations at the beginning of the processing interval; while instability ($a \ge 0$) or driving noise ($c \neq 0$) causes the effort to be placed at the

end of the interval. When $a < 0$ and $c \neq 0$ there exists a trade off and part of the effort is placed at each end of the processing interval.

Solution of Case 2

The scalar version of case 2 presents a more interesting problem solution for the possibility of singular arcs in the optimal trajectories is not ruled out. Recalling the formulation of the problem, minimize

$$J_2(u) = \int_0^T fp(t)+\alpha u dt \tag{4.11}$$

with respect to u, subject to

$$\dot{p} = 2ap+c-up^2d \tag{4.12}$$

with p(0) given. The Maximum Principle yields an equivalent boundary value problem. Determine the optimal allocation function u^* according to

$$u = \begin{cases} 1 & \psi(p,z) > 0 \\ 0 & \psi(p,z) < 0 \\ ? & \psi(p,z) = 0 \end{cases}$$

subject to

$$\begin{aligned} \psi(p,z) &= zp^2+\alpha \\ \dot{p} &= 2ap+c-up^2d \\ \dot{z} &= 2(pdu-a)z+f. \end{aligned} \tag{4.14}$$

This case as the previous problem is divided into two subproblems according to the sign of a. By investigating the interrelations of the critical surfaces in the pz plane and the direction in which they are crossed by trajectories the form of the optimal allocation is deduced. Once the form is known simple calculations give the switching times.

$a \geq 0$. The situation for this case is illustrated by figure 2. The switching curve $\psi=0$ separates the quadrant into the regions where u=0 and u=1 as shown. Considering the inner product of the gradient of the switching curve with the time derivatives of the trajectories

$$\langle(\dot{p},\dot{z}); \nabla\psi(p,z)\rangle = 2pdz(ap+c+pf/2z), \tag{4.15}$$

determines that these are perpendicular at only one point (p_s,z_s). This point is the only candidate for a singular arc of the optimal trajectory. To the left of the point trajectories cross into the u=1 region and to the right into the u=0 region. The curve labeled Δ is the locus of points (p,z) such that $\dot{(p,z)}=0$ for allowable u. As this line intersects the switching curve at $(\dot{p}_s,\dot{z}_s)$ the possibility of a singular portion of the trajectory remains viable.

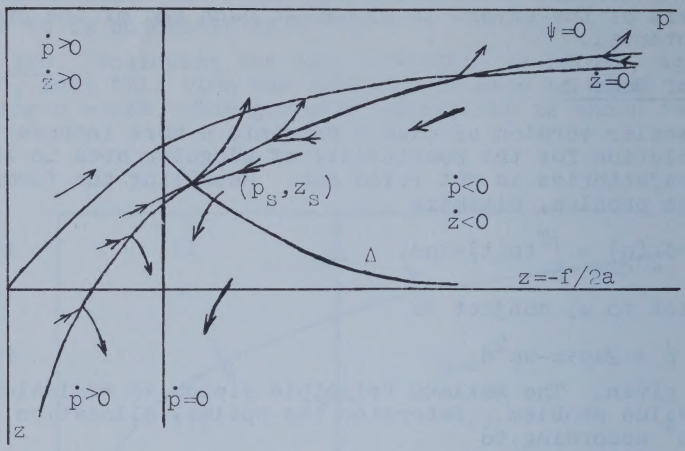


Figure 2. The pz plane for case 2 with $a \geq 0$.

Next, insert the curves $p=0$ and $\dot{z}=0$ where $u=1$ and note the direction of the time derivative of the trajectories (these directions are indicated by arrows). Remembering that $z(T)=0$, the direction of the trajectories rules out the possibility of an optimal trajectory crossing the switching boundary into the region $u=1$. Further, a trajectory can enter the singular point only once. Thus if u_s denotes the value of u that maintains the trajectory at the singular point, u must be one of the sequences

$$\begin{array}{l} 1,0 \\ 1,u_s,0 \\ 0,u_s,0 \\ u_s,0 \\ 0. \end{array} \tag{4.16}$$

Since the terminal condition on z is known and $\dot{z}$ is independent of p for $u=0$, the duration of a $u=0$ arc following a singular arc is computed as

$$\tau_0 = -\ln[f/(2az_s+f)]/2a. \tag{4.17}$$

Further the duration of $u=0$ or $u=1$ arcs proceeding a u_s arc is also easy to compute

$$\theta_0 = \ln[(2ap_s+c)/(2ap(0)+c)]/2a \tag{4.18}$$

and θ_1 is the solution of

$$p_s = \left[c/g+p(0)(\coth(g\theta_1)+a/g)\right]/\left[\coth(g\theta_1)-a/g+\frac{dp(0)}{g}\right] \tag{4.19}$$

where $g = -(a^2+dc)^{\frac{1}{2}}$ By comparing these intervals with T the type of allocation actually present in a given problem is easily ascertained. For example, if $\tau_0+\theta_0>T$ the sequence $0,u_s,0$ is impossible.

$a\leq 0$. As figure 3 shows this case is very similar to the previous part of the problem. A primary difference is the other set of lines which satisfy $\dot{p}=0$ and $\dot{z}=0$ when $u=0$. Because of the further restriction that they create the possibility of a singular arc can be eliminated immediately if (p_s,z_s) does not lie in the area formed by the intersection of the four curves. Otherwise, the form is the solution is the same as for $a>0$.

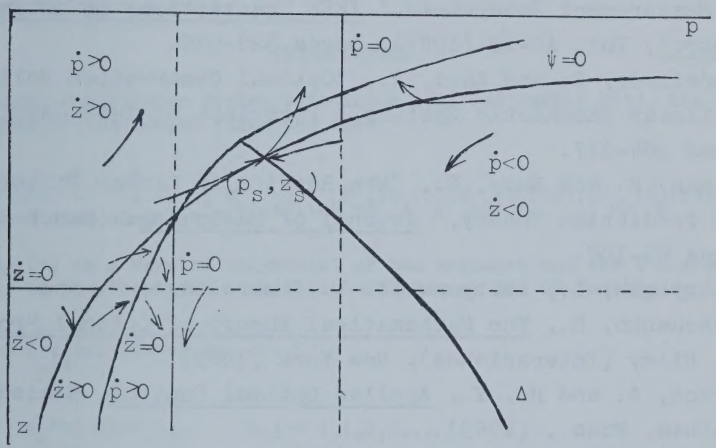


Figure 3. The pz plane for case 2 with $a<0$.

V. CONCLUSION

An optimal allocation problem which yields a nonlinear optimal control problem has been formulated and solved for the scalar case. It is logical to ask if the techniques used to solve the scalar problem can be used in higher dimensions. As the method employed in the paper was strongly geometric, extensions seem unlikely. However, the Ricatti equation has very interesting monotonicity properities under special circumstances. Hence, one might hope to derive analytic conditions for which the optimal allocation contains at most one switch. In any case, a general solution to the problem in higher dimensions seem unattainable.

REFERENCES

1. Athans, M. and Schweppe, F., "Optimal Waveform Design via Control Theoretic Concepts," Information and Control, Vol. 10 (1967), pages 335-377.
2. Schweppe, F. "State Space Evaluation of the Bhattacharyya Distance between Two Gaussian Processes," Information and Control, Vol. 11 (1967), pages 352-372.
3. Meier, L., Peschon, J., and Dressler, R. "Optimal Control of Measurement Subsystems," IEEE Transactions on Automatic Control, Vol. AC-12 (1967), pages 528-536.
4. Vandelinde, D. and Lavi, A., "Optimal Observation Policies in Linear Stochastic Systems," Preprints of 1968 JACC, pages 904-917.
5. Kalman, R. and Bucy, R., "New Results in Linear Filtering and Prediction Theory," Journal of Basic Eng. (March 1961) pages 95-108.
6. Pontryagin, L., Boltyanskii, V. Gamkrelidge, R. and Mishchenko, E., The Mathematical Theory of Optimal Processes, Wiley (Interscience), New York (1962)
7. Bryson, A. and Ho, Y., Applied Optimal Control, Blaisdell, Waltham, Mass., (1969).

IMPLEMENTATION OF THE WIENER MODEL*

R. J. KUHLER and R. H. SEACAT
Texas Tech University
Lubbock, Texas 79414

ABSTRACT

The expansion of an arbitrary nonlinear functional into an infinite series of orthogonal functions was first examined in detail by Wiener and has recently been applied to various mathematical systems. In this investigation an inertial quality accelerometer, which exhibits various nonlinear effects, was modeled using Wiener's methods. Error sensitivity methods to determine coefficients of the model are compared to the original Wiener technique.

INTRODUCTION

Given a nonlinear system with input x(t) and output y(t), the Wiener expansion of the output takes the form

$$y(t) = \lim_{N\to\infty} \sum_{\alpha 1=0}^{\infty} \sum_{\alpha 2=0}^{\infty} \cdots \sum_{\alpha N=0}^{\infty} C_{\alpha 1,\alpha 2 \cdots \alpha N} H_{\alpha 1}(V_1(t)) H_{\alpha 2}(V_2(t)) \cdots H_{\alpha N}(V_N(t)) \qquad (1)$$

$H_{\alpha i}(\cdot)$ is a Hermite polynomial of the argument and the V's are convolutions of the input and a suitable orthonormal set of functions, i.e.

$$V_j(t) = \int_0^t \phi_j(\tau)\, x(t-\tau)\, d\tau \qquad (2)$$

$$\int_0^{\infty} \phi_j(\tau)\, \phi_k(\tau)\, d\tau = \delta_{jk} \qquad k,j = 0,1,2,3\ldots \qquad (3)$$

Under an input of white Gaussian noise the V's are statistically independent and the joint probability density functions of the V's provide the condition for orthogonality of the Hermite polynomials. Envoking the ergodic theorem gives the weighting of the terms in the model as

$$C_{\alpha 1 \alpha 2 \ldots \alpha N} = \left\langle y(t)\, H_{\alpha 1}(V_1(t))\, H_{\alpha 2}(V_2(t)) \cdots H_{\alpha N}(V_N(t)) \right\rangle \qquad (4)$$

IMPLEMENTATION

To employ the identification procedure an inertial quality accelerometer was placed under an acceleration input approaching the white Gaussian properties desired. The spectrum of the input noise was shaped using sectioned filters whose input was a reference acceleration signal. The output of the filters were feed back to flatten the spectrum of the input signal. The input and output signals were digitized and digital Laguerre filters were used to generate the basis functions for the orthonormal expansion.

The model's rms error versus the number of terms taken in the expansion were plotted for three different methods of series expansion. The first method determined the coefficients from rms error minimization. The second method utilizes the time averaging indicated by equation (4) and

each term is removed from the system output as it is determined. The third method consisted of removing a linear second order system before application of the second method. The results of each technique are shown in Figure 1. In each case 14 linear terms were removed first then the cross terms beginning with lower order terms as indicated by subscripts in equation (4) were removed.

CONCLUSION AND OBSERVATIONS

For the example studied no apparent advantage is obtained by selecting coefficients on an rms error sensitivity basis as the Wiener method gives comparable results with less computations. The application of the Wiener identification technique to residual error between a model and the physical system is shown by method 3 to be promising. The fact that the small signal model performed poorly under the noise environment is of no consequence.

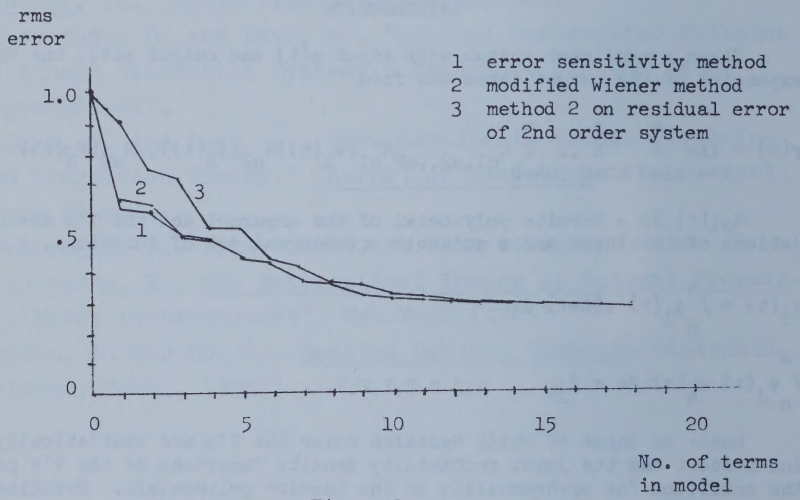


Figure 1

* This work supported in part by the U. S. Air Force, contract number AF 29(600-5590)

REFERENCES

1. Arnold, C. R., "Identification of Systems in Terms of the Wiener Model", MIT Lab T.N. 1967-34, Lexington, Mass. (August 1967).

2. Wiener, N., Nonlinear Problems in Random Theory, John Wiley and Sons, Inc., New York, 1958.

M-ARY CLASSIFICATION -- THE CASE OF UNKNOWN A PRIORI INFORMATION

YEN-BIN CHEN
IBM Systems Development Division
Kingston, New York

ABSTRACT

The problem of M-ary signal classification in a statistically undefined noise background is analyzed in situations in which no a priori information for the signal is available. It is well-known that, in such an environment, the best solution is the minimax procedure -- when this procedure exists. The formulation for the minimax solution introduced here shows that it may not be possible to obtain such a solution analytically. An alternative is presented which, though less efficient, can be implemented by using simple counters and/or logic circuitry.

I. Introduction

Generally, when the a priori information of the signal was known, the problem of classifying M-ary signals in additive noise was resolved some time ago. However, little work has been performed in cases in which the noise statistic is nonparametric. In fact, to the best of the author's knowledge, no attempt has been made to obtain an optimum solution if, in addition, there is no a priori information. Similarly, the intent of this paper is not to resolve the problem through finding the optimum solution; instead, a minimax formulation of the problem is presented.

Minimax formulation for M-ary classification in gaussian noise (when the covariance matrices are the same) has been given by Anderson [1]; frequently, however, the solution comes only through the use of trial-and-error methods. If the noise distribution is unknown, it will be shown in this paper how, at least in theory, a minimax procedure could be formulated. However, in such cases, the difficulty in obtaining a solution analytically may make the minimax approach virtually untenable.

Because of the difficulties inherent in a minimax solution, an alternative proposal will be made. Although the proposal constitutes an ad-hoc procedure (like many other nonparametric detectors), it is suggested, however, by the symmetrical theorem that it is not unreasonable. This procedure assumes no knowledge about the noise distribution. It only requires that the class of noise under consideration be symmetrical about zero. Because the noise statistic considered is so broad that it can no longer be characterized by a finite number of parameters, the scheme is essentially nonparametric. In formulating the test statistic, amplitude information was ignored; instead, polarity information was used. For this reason, implementation is very simple.

II. Minimax Formulation

Since the noise distribution is unknown, a Bayes likelihood ratio cannot be derived. We can, however, neglect the sample amplitude information and form a new set of observable random variables based on cell counts which are defined as the number of samples falling in each of the $(M + 1)$ cells partitioned by the M-signals. It is clear that these cell counts are random variables multinomially distributed, and a Bayes likelihood ratio is

derived as follows.*

Let $\mu_1, \mu_2, \ldots, \mu_M$ be the M signals where $\mu_1 < \mu_2 < \ldots < \mu_M$ with respective known a priori information $P_1, P_2, \ldots, P_M$. $\chi_1, \chi_2, \ldots, \chi_n$ are the observable random samples (a mixture of signal plus noise) from continuous c.d.f. $F(\chi)$. Let us partition the sample space into $(M + 1)$ cells by the M signal such that cell 0 is in between $-\infty$ and μ_1, cell 1 is in between μ_1 and μ_2, and so on. Finally, cell M is in between μ_M and $+\infty$. Furthermore, let the corresponding cell counts be $n_0, n_1, \ldots, n_M$ respectively. Since $n_0 + n_1 + \ldots + n_M = n$, it is apparent that $n_0, n_1, \ldots n_M$ are multimonially distributed as

$$\binom{n}{n_0, \ldots n_M} \prod_{i=0}^{M} q_{k(i)}^{n_i} \qquad \begin{matrix} i = 0, 1, \ldots, M \\ k = 1, 2, \ldots, M \end{matrix}$$

where $q_k(i) = F_k(\mu_{i+1}) - F_k(\mu_i)$

and $F_k(\mu) = P(\chi < \mu \mid \mu_k)$

According to Bayes criterion, the best possible decision rule in this reduced sample space is the likelihood ratio functional and can be easily shown as

$$\Lambda_{kj}(n_i) = \underline{N} \cdot \underline{\theta}_{kj} \qquad (1)$$

which is known as the Bayes discriminate function and $N = (n_0, n_1, \ldots, n_M)$ is a random vector,

$$\underline{\theta}_{kj} = \left(\log \frac{q_{k(0)}}{q_{j(0)}}, \log \frac{q_{k(1)}}{q_{j(1)}}, \ldots, \log \frac{q_{k(M)}}{q_{j(M)}}\right)$$

is a vector of nuisance parameters. In the formulation described below, we shall assume that these parameters were known, and the theory applies to large sample case when the nuisance parameters are replaced by their estimates [4].

If the priori probabilities $P_1, P_2, \ldots, P_M$ are known, the Bayes solution will classify the signal as μ_k by those n_i's in region R_k satisfying

$$R_k : \quad \Lambda_{kj}(n_i) > \log \frac{P_j}{P_k} \qquad j = 1, 2, \ldots, M, \; j \neq k \qquad (2)$$

since the procedure is Bayes, the regions of classification $R_1, R_2, \ldots, R_M$ defined by (2) would minimize the expected cost. In the case of n_0 a priori probability is known, the region R_k is defined by the inequalities

$$\Lambda_{kj}(n_i) \geq C_k - C_j \qquad j = 1, \ldots, M, \; j \neq k \qquad (3)$$

The set of regions form the class of admissible procedures. For the minimax solution, these constants are determined so that all the probabilities of correct classification $P(K|\mu_k, R)$ are equal. The difficult for a solution is that of determining these probabilities. Consider the asymptotic multinormal distribution of $\Lambda_{kj}(n_i)$, the probability of correct classification can be written as

*For a detailed discussion on the derivation and application to M-ary detection, see reference [4].

$$P(k|\ \mu_k,\ R) = \int_{C_k}^{\infty} \cdots \int_{-C_M}^{\infty} f_k\, d\lambda_{k1} \cdots d\lambda_{k,\,k-1}\, d\lambda_{k,\,k+1} \cdots d\lambda_{kM} \quad (4)$$

where f_k is the asymptotic multinormal density of $\Lambda_{kj}(n_i)$ for $j = 1,2, \ldots, M,\ j \neq k$.

Theorem If the noise has continuous c.d.f. and the cost of misclassification are equal, the regions of classification $R_1, R_2, \ldots, R_M$ that minimze the maximum conditional expected loss are defined by (3) where $\Lambda_{kj}(n_i)$ is given by (1). The constants C are determined so that the integral (4) are equal.

It is seen that minimax procedure can be formulated in theory. Unfortunately, we have not found possible any direct way to obtain the C's because the covariance matrices of the asymptotic multinormal distribution under different signal hypotheses are different and that resulted in quadratic discriminate function rather than linear. We do not know at present the distribution of a quadratic discriminate function.

III. The Symmetrical Detector

Let us consider the following multiple hypotheses testing problem.

Take random samples $\chi_1, \chi_2, \ldots,$ n from a continuous c.d.f. $F(\chi|\mu)$ with mean μ. The multiple hypotheses (shift only) are

$H_1 = \mu = \mu_1$

$H_2 = \mu = \mu_2$

$\vdots$

$H_M = \mu = \mu_M$

Furthermore, if we consider the class of distributions that are symmetrical about zero, then we can formulate the following theorem.

The Symmetrical Theorem

Let $\chi_1, \chi_2, \ldots, \chi_n$ be defined as above. Form the set of statistics D_j as

$$D_j = \left| \sum_{i=1}^{n} [\, U(\chi_i - \mu_j) - U(\mu_j - \chi_i)] \right| \qquad j = 1, 2, \ldots, M$$

where $U(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \end{cases}$

D_j measures the degree of symmetry of the samples about mean μ_j, then

$F(D_k) < E(D_j)$ for all $j \neq k$

if hypothesis H_k (signal μ_k sent) is true.

Proof: The probability function of D_j is shown in appendix to be given by

$$P[D_j = d \mid \mu_\ell] = \binom{n}{\frac{n-d}{2}} p_{j\ell}^{\frac{n-d}{2}} (1 - p_{j\ell})^{\frac{n+d}{2}} \left[\left(\frac{p_{j\ell}}{1 - p_{j\ell}} \right)^{d} + 1 \right] \cdots \quad (5)$$

where $P[D_j = d \mid \mu_\ell]$ is the probability of D_j under hypothesis H_ℓ with mean μ_ℓ, and

$$P_{j\ell} = P[\chi_i > \mu_j \mid \mu_\ell]$$

Assume $n=2r+1$ and let us make the following transformation

$$Y_j = \frac{n - D_j}{2}$$

Then $$P[Y_j = y \mid \mu_\ell] = \binom{n}{y} p_{j\ell}^{y} (1-p_{j\ell})^{n-y} \left[\left(\frac{p_{j\ell}}{1-p_{j\ell}} \right)^{n-2y} + 1 \right] \quad \ldots (6)$$

and $D_j = 1, 3, 5, \ldots\ldots, n=2r+1$, $Y_j = 0, 1, 2, \ldots\ldots (n-1)/2 = r$

It is clear that

$$\sum_{y=0}^{r} \binom{n}{y} p_{j\ell}^{y} (1-p_{j\ell})^{n-y} \left[\left(\frac{p_{j\ell}}{1-p_{j\ell}} \right)^{n-2y} + 1 \right] = 1 \quad \ldots (7)$$

Taking expected value of Y_j under H_ℓ

$$E(Y_j \mid \mu_\ell) = \sum_{y=0}^{r} y \binom{n}{y} p_{j\ell}^{y} (1-p_{j\ell})^{n-y} \left[\left(\frac{yj1}{1-p_{j\ell}} \right)^{n-2y} + 1 \right]$$

$$= n(n-1) p_{j\ell} (1-p_{j\ell}) \sum_{y=1}^{r} \binom{n-2}{y-1} p_{j\ell}^{y-1} (1-p_{j\ell})^{n-2-(Y-1)} \left[\left(\frac{p_{j\ell}}{1-p_{j\ell}} \right)^{n-2-(y-1)} + 1 \right]$$

Let $n - 2 = m = 2s + 1$. $y - 1 = Z$

$$\therefore \quad S = r - 1 = \frac{m-1}{2}$$

$$E[\bar{Y}_j \mid \mu_\ell] = n(n-1) p_{j\ell} (1-p_{j\ell}) \sum_{Z=0}^{s} \binom{m}{Z} p_{j\ell}^{Z} (1-p_{j\ell})^{m-Z} \left[\left(\frac{p_{j\ell}}{1-p_{j\ell}} \right)^{m-2Z} + 1 \right]$$

According to Eq. (7), we have

$$\sum_{Z=0}^{s} \binom{m}{Z} p_{j\ell}^{Z} (1-p_{j\ell})^{m-2} \left[\left(\frac{p_{j\ell}}{1-p_{j\ell}} \right)^{m-2Z} + 1 \right] = 1$$

$$\therefore \quad E(Y_{j\ell} \mid \mu_\ell) = n(n-1) p_{j\ell} (1-p_{j\ell})$$

Since $0 \leq p_{j\ell} \leq 1$

$E(Y_j / \mu_\ell)$ reaches its maximum

at $p_{j\ell} = 1/2$.

But $p_{j\ell} = P[\chi_i > \mu_j \mid \mu\ell] = 1/2$ if $j=\ell$

i.e. $E(Y_\ell \mid \mu\ell) > E(Y_j/\mu\ell)$ all $j\neq\ell$

$E(D_\ell /\mu\ell) < E(D_j/\mu\ell)$ all $j\neq\ell$

For $n=2r$

$D_j = o, 2, 4, \ldots\ldots, n=2r$

$Y_j = 0, 1, 2, \ldots\ldots, n/2=r$

and the proof is identical.

Q. E. D.

Decision Procedure

Assume that we have M-deterministic signals $\mu_1, \mu_2, \ldots \mu_M$ with unknown a priori probabilities of occurrence. Every T seconds one of the M possible signals is transmitted and being corrupted by statistically undefined noise. At the receiver side we propose the following procedure to decide that within the T seconds, which one of the M signals has been transmitted. The procedure is

(a) Observe random samples $\chi_1, \chi_2, \ldots, \chi_n$ of size n.

(b) Compute for j = 1, 2, ..., M

$$D_j = \left| \sum_{i=1}^{n} [U(\chi_i - \mu_j) - U(\mu_j - \mu_i)] \right|$$

(c) Choose μ_k if $D_k = \min(D_1, D_2, \ldots, D_M)$. In case of tie, randomized procedure is used.

Discussion

The probability of error for binary classification in the case of equal a priori probability in both gaussian and Laplacian noises are simulated using Monte Carlo method as shown in Figure (1). Unfortunately, there is no counterpart that we can compare with. Nevertheless, comparing with what the parametric cross-correlator would do when a priori probabilities are known to be equally likely, even in this case, the symmetrical detector is not grossly inferior. In addition no analog processing of numerical values are required in implementation which depends only on polarity counts.

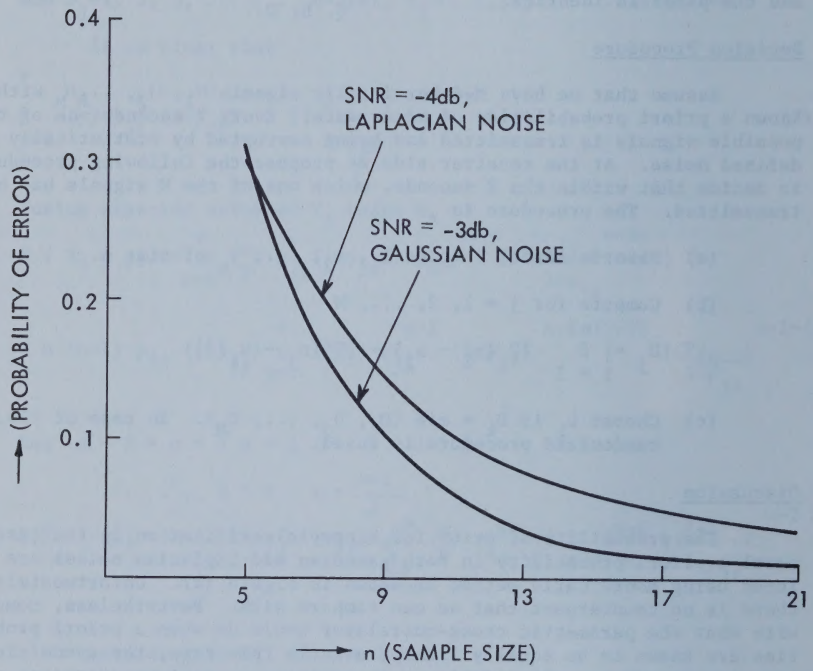


FIG. (1) PROBABILITY OF ERROR VS SAMPLE SIZE

APPENDIX

Probability Function of D-statistic

We shall now derive the probability function P[D=d] as stated in the symmetrical theorem for the symmetrical detector as follows:

The statistic D has been defined as

$$D = \left| \sum_{i=1}^{n} [U(\chi_i - \mu) - U(\mu - \chi_i)] \right|$$

$$= |2k - n|$$

where k is the number of χ_i's $> \mu$ in the sample observations $\chi_1, \chi_2, \ldots, \chi_n$. It is clear that D is the absolute value of Binomial r.v., say W, i.e., D = |W|. It is clear that with $P(\chi_i > \mu) = p$

$$P[W=2K-n] = \binom{n}{k} p^k (1-P)^{n-k}$$

= P (probability that there are k of χ_i's which are greater than μ)

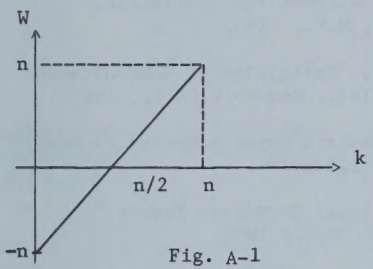


Fig. A-1

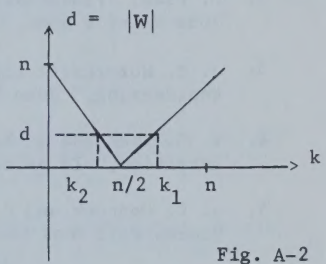


Fig. A-2

From Figs. A-1 and A-2, we see that

$$P[D=d] = P[|W| = |2k-n|]$$

$$= P[D = |W|]$$

$$= P[2k_1 - n = d] + P[2k_2 - n = -d]$$

$$= P[k_1 = \frac{n+d}{2}] + P[k_2 = \frac{n-d}{2}]$$

$$= \binom{n}{k_1} P^{k_1} (1-P)^{n-k_1} + \binom{n}{k_2} P^{k_2} (1-P)^{n-k_2}$$

Since $k_1 = \frac{n}{2} + \frac{d}{2} = k_2 + d$

$$k_2 = \frac{n-d}{2}$$

$$P[D=d]$$

$$= \binom{n}{k_2+d} P^{k_2+d} (1-P)^{n-k_2-d} + \binom{n}{k_2} P^{k_2} (1-P)^{n-k_2}$$

$$= P^{k_2} (1-P)^{n-k_2} [(k_2^n + d) P^d (1-P)^{-d} + (k_2^n)]$$

$$= \binom{n}{\frac{n-d}{2}} P^{\frac{n-d}{2}} (1-P)^{\frac{n+d}{2}} \left[\left(\frac{P}{1-p} \right)^d + 1 \right]$$

References

1. T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," John Wiley & Sons, Inc., New York, N.Y., 1958.

2. M. Fisz, "Probability Theory and Mathematical Statistics," John Wiley & Sons, Inc., New York, N.Y., 1963.

3. J. M. Wozercraft and I. M. Jacobs, "Principles of Communication Engineering," John Wiley & Sons, Inc., New York, N.Y., 1965.

4. Y. B. Chen and A. Nadas, "Nonparametric Bayes Approach to M-ary Detection," To be submitted for publication.

5. J. C. Hancock and P. A. Wintz, "Signal Detection Theory," McGraw-Hill Book Company, New York, N.Y., 1966.

6. D.A.S. Fraser, "Nonparametric Method in Statistics," John Wiley & Sons, Inc., New York, N.Y., 1958.

ON THE FORMULATION OF DICHOTOMOUS FLOWGRAPHS FOR NETWORKS HAVING EXCESS ELEMENTS

E. M. RIPS
Newark College of Engineering, Newark, New Jersey

ABSTRACT

The dichotomous flowgraph used in the NASAP program represents relations between tree-branch voltages and chord currents for a proper tree. Capacitive loops or inductive outsets may be dealt with by modifying the flowgraph by means of an excess-variable reduction algorithm.

CIRCUIT TOPOLOGICAL RELATIONSHIPS

For a passive network with no excess elements, tree-branch voltages, chord currents and network parameters are related according to Eq. (1).

$$\begin{bmatrix} sCV_c \\ GV_G \\ sLI_L \\ RI_R \end{bmatrix} = \begin{bmatrix} 0 & B_f \\ -B_f^t & 0 \end{bmatrix} \begin{bmatrix} V \\ I \end{bmatrix} + \begin{bmatrix} J \\ E \end{bmatrix} \qquad (1)$$

where V and I are tree-branch voltage vectors and chord current vectors respectively, J and E are vectors representing independent current or voltage sources, and B_f is a fundamental circuit matrix.

If, as in the NASAP program, there is a single independent source, we can partition Eq. (1) to yield Eqs. (2). For this discussion we are taking the single source to be a voltage generator, e.

$$\begin{bmatrix} sC \\ G \end{bmatrix} [V] = [B_f][I] \qquad (2a)$$

$$\begin{bmatrix} sL \\ R \end{bmatrix} [I] = [-B_f]^t [V] + e \qquad (2b)$$

In Eqs. (2) the matrices $\begin{bmatrix} sC \\ G \end{bmatrix}$ and $\begin{bmatrix} sL \\ R \end{bmatrix}$ are square and conformable with $[V]$ and $[I]$, respectively.

CONSTRUCTION OF FLOWGRAPH

Each equation embedded in the matrix collection given by Eqs. (2) can be divided through by the appropriate coefficient of the current or voltage that appears on the left-hand side. Eqs. (2) can then be represented by the dichotomous flowgraph shown in Fig. 1.

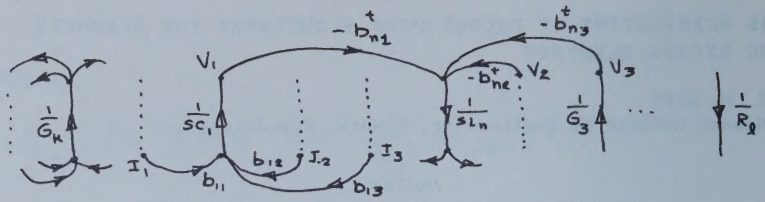


Fig. 1. Basic Features of Dichotomous Flowgraph.

Of course, transmittances b_{ij} have values of +1, -1 or zero. Active networks are treated in the same fashion, except that non-symmetrical non-zero elements may arise to complicate Eqs. (1) and (2).

The current Li_x in an excess inductor L_x can be expressed in terms of the currents in the inductive cutset to which it belongs. In similar fashion the voltage v_{ey} across an excess capacitor C_y can be expressed in terms of the remaining capacitive voltages of its loop.

Since non-excess inductors belong to the set of chords while non-excess capacitors belong to the set of tree branches, Eq. (1) can be modified to yield

$$\begin{bmatrix} sCV_c \\ i_{Lx} \\ GV_g \\ v_{cy} \\ sLI_L \\ RI_R \end{bmatrix} = \begin{bmatrix} P & B_f \\ -B_f^t & Q \end{bmatrix} \begin{bmatrix} v_c \\ v_{Lx} \\ V_g \\ i_{cy} \\ I_L \\ I_R \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ e \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (3)$$

The null matrices that appeared in Eq. (1) are replaced in Eq. (3) by P and Q to account for controlled sources.

Since $v_{Lx} = sL_x i_{Lx}$ and $I_{cy} = sC_y v_{cy}$, both v_{Lx} and i_{cy} can be eliminated and the order of Eq. (3) reduced accordingly. The price of this elimination of variables is the modification of the L and C matrices. In modified form, these matrices will, in general, no longer be diagonal. Let the modified matrices be designated L' and C', respectively.

Incorporating these modified matrices into Eq. (3) yields

$$\begin{bmatrix} sC'V_c' \\ GV_g \\ sL'I_L' \\ RI_R \end{bmatrix} = \begin{bmatrix} P' & B_f' \\ (-B_f^t)' & Q' \end{bmatrix} \begin{bmatrix} V_c' \\ V_g \\ I_L' \\ I_R \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ e \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (4)$$

where the primes indicate deletion of appropriate rows and columns and modification of the remaining elements.

From Eq. (4) the modified dichotomous flowgraph can be constructed.

NASAP ANALYSIS OF NONLINEAR DC CIRCUITS*

K. Haag
Illinois Institute of Technology
Chicago, Illinois

ABSTRACT

This paper presents a method for using a linear circuit analysis program to analyze nonlinear dc circuits. An option on the NASAP program is used which results in real variable calculation of the transfer function. Auxilliary circuits which model the nonlinearity are developed and used for transistor bias circuit analysis and power supply regulator analysis.

INTRODUCTION

The output obtained when using the NASAP program is a transfer function in terms of the variable s. By making a slight modification in the standard NASAP program the evaluation of the transfer function is made in terms of a real variable, σ, rather than the imaginary variable, $j\omega$. Using this additional degree of freedom, nonlinearities in dc circuits can be modeled with an auxilliary circuit and the NASAP analysis produces a tabular and graphical output for the nonlinear circuit.

This paper develops the basic procedure for analyzing four classes of nonlinear dependencies: a) Nonlinear controlled source dependent on current or voltage in the circuit, b) Nonlinear controlled source dependent on a parameter external to the circuit such as temperature, c) Nonlinear resistor dependent on current or voltage associated with its terminals, and d) Nonlinear resistor dependent on a parameter external to the circuit.

A method for obtaining models for nonlinear elements is presented and is used to establish a model for an exponential nonlinearity. This model is used in the analysis of transistor bias circuits.

The paper also includes analysis procedures for power supply regulator circuits using the real s evaluation option. This analysis produces results related to the load and line regulation properties of the regulator.

MODELING CIRCUIT ELEMENT VARIATIONS

Linear and nonlinear circuit element variations in a resistive circuit can be simulated with the use of the SIGMA option of the NASAP program which results in a real variable evaluation of the transfer function. Consider a

---

*This work was supported by the National Aeronautics and Space Administration under Contract No. 12-2064.

variable circuit element A(x) where A may be a resistor, a current source, or a voltage source and x may be a current, a voltage, or some external parameter such as temperature. The function A(x) is approximated by a ratio of polynominals using either approximation methods or trial and error.

$$A(x) = \frac{P(x)}{Q(x)} \tag{1}$$

The variable x is replaced by s and a circuit is synthesized which has a transfer function or an input immitance given by P(s)/Q(s). When the SIGMA option is used, this circuit will simulate the variation of the circuit element A(x) if the range of s is the same as the expected range of x. It should be noted that since the circuit is used for simulation purposes only, it does not have to be physically realizable, i.e., controlled sources or negative circuit elements may be used.

The method for using this model in computer analysis of a circuit depends on the circuit element and variable which A and x represent. In Cases I and II of the discussion which follows it is assumed that P(s)/Q(s) is realized as an input impedance. If the realization is an input admittance or a transfer function, relatively minor modifications are required.

Case I

A is a controlled voltage or a current source and x is a voltage or a current in the circuit.

The dependency function A(s) is synthesized as an impedance and a current generator of value unity is connected to the impedance. This auxilliary network is hinged to the network which is to be analyzed. The nonlinear dependent source in the network is replaced by a source which is dependent on the voltage across the auxilliary network. The transfer function x/independent source is specified and the SIGMA option is used. The solution for the circuit is obtained from the tabular printout or the graphical plot as the value where x = s. If x is not the desired output variable, the circuit is recoded with A(x) replaced by its known value from the above solution. The auxilliary network is eliminated and the desired output variable is specified.

This analysis is applicable when the circuit contains no more than one variable source.

To illustrate the method, consider the circuit of Figure 1. Let R = 2 and assume that V is a function of I_1 and is adequately approximated by

$$V(I_1) = \frac{I_1^2 + I_1 - 5}{I_1 + 1} \qquad 0 \leq I_1 \leq 1 \tag{2}$$

A pencil and paper solution results in

$$I_1 = .921 \; ; \; V_o = .0795 \tag{3}$$

For computer analysis an impedance is synthesized for which

$$Z(s) = \frac{s^2 + s - 5}{s + 1} = s - \frac{5}{s + 1} \tag{4}$$

A circuit realization of this impedance is shown in Figure 2. The auxilliary network is hinged to the basic circuit with R2 = 2, I1 = IR7 and V2 = VR6. Note that the R5,R6 resistance combination is used for voltage measurement only and could be replaced by a single, large valued resistance. The transfer function IR1/VV1 is specified and a frequency of .915 to.925 in steps of .001 is specified. The frequency range is chosen on the basis of the expected "ballpark" value of the current IR1. If a good estimate of IR1 cannot be made, a larger frequency range is specified with larger increments and more than one pass on the computer may be necessary to obtain IR1 to the desired degree of accuracy.

The computer analysis results in

$$\frac{IR1}{VV1} = -.25 \; \frac{-7 - s + s^2}{1 + s} \tag{5}$$

and the printout table gives IR1 ≃ s for s = .921. The value of V2 is then obtained as V2 = -1.682. The basic circuit is then coded using this value of V2 and the result is VR3/VV1 = .0795.

Case II

A is a controlled voltage or current source and x is an external parameter which is known to vary over a specified range.

In this case, the controlled source is generally dependent on a current or voltage in the circuit (denoted by y in this discussion) and the dependency factor is a function of an external parameter x. The $h_{FE}I_B$ generator in the transistor dc equivalent circuit where h_{FE} is temperature dependent is an example of this situation. The circuit is coded with yA(x) replaced by a generator, V(s), which is dependent on the voltage across the auxilliary network. The desired transfer function is specified and the SIGMA option is used. The computer solution gives the relation between the output variable and x in functional (replace s by x), tabular and graphical forms. This method is applicable for any number of variable sources in the circuit which depend on the same external parameter x.

The circuit of Figure 1 is used to illustrate the procedure with R = 2 and V dependent on I_1 and temperature according to

$$V = I_1 \frac{T^2 + T - 5}{T + 1} \qquad 10 \leq T \leq 50 \tag{6}$$

A hand solution results in

$$V_o = \frac{T^2 + 3T - 3}{T^2 + 5T - 1} \tag{7}$$

For computer analysis the nonlinear dependency has the same functional form as in Case I and therefore Z(s) of Figure 2 simulates the temperature dependency. The auxilliary network is hinged to the basic circuit with I1 = IR1 and V2 = VR6. The transfer function VR3/VV1 is requested and a frequency range of 10 to 50 in steps of 1 is specified. The result is

$$\frac{VR3}{VV1} = \frac{-3 + 3s + s^2}{-1 + 5s + s^2} \tag{8}$$

A table is printed out and a graph of the function over the range s = 10 to s = 50 is plotted.

Case III

A is a resistor and x is the voltage across or the current through the resistor.

The circuit is coded with R(x) replaced by Z(s). The transfer function x/independent source is specified and the SIGMA option is used. The solution for the circuit is obtained from the tabular printout or the graphical plot as the value where x = s. If x is not the desired output variable, the circuit is recoded with R(x) replaced by its known value and the desired transfer function is specified. This analysis is applicable when the circuit contains no more than one nonlinear resistance.

In the circuit of Figure 1, let V = -2 and R be dependent on I_2.

$$R(I_2) = \frac{I_2^2 + I_2 - 5}{I_2 + 1} \qquad 2 \leq I_2 \leq 4 \tag{9}$$

The solution is

$$I_2 = 2.155; \quad V_o = -.771 \tag{10}$$

For computer analysis, the basic circuit of Figure 2 is used with R2 replaced by Z(s) and V2 = -2 IR7. The transfer function IL1/VV1 is requested and a frequency range of 2.15 to 2.16 in steps of .001 is specified. As in Case I, it may require more than one pass on the computer to determine the proper frequency range and increment which will result in the desired degree of accuracy for IL1. The computer analysis gives

$$\frac{IL1}{VV1} = 2.67 \frac{1 + s}{-4.333 + 1.667s + s^2} \tag{11}$$

for which IL1 ≃ s at s = 2.155. The value of $R(I_2)$ is evaluated as $R(I_2)$ = .5702. The basic circuit is recoded using this value of R and the above value for V. The analysis results in VR3/VV1 = -.771.

Case IV

A is a resistor and x is an external parameter which is known to vary over a specified range.

The circuit is coded with R(x) replaced by Z(s) and the desired transfer function is specified. The computer solution gives the relation between the output variable and x. This method is applicable for any number of variable resistances in the circuit which depend on the same external parameter x.

As an example, let V = 2 and R be a function of temperature in the circuit of Figure 1.

$$R(T) = \frac{T^2 + T - 5}{T + 1} \qquad 10 \leq T \leq 50 \tag{12}$$

A hand solution gives

$$V_o = .667 \frac{T^2 + 3T - 3}{T^2 + 1.667T - 4.333} \tag{13}$$

In the computer analysis, the basic circuit of Figure 2 is used with R2 replaced by Z(s) and V2 = 2IR7. The frequency range 10 to 50 in steps of 1 is specified and the computer output is

$$\frac{VR3}{VV1} = .667 \frac{-3 + 3s + s^2}{-4.333 + 1.667s + s^2} \tag{14}$$

A table is printed out and a graph of the function over the range s = 10 to s = 50 is plotted.

MODELING TEMPERATURE VARIATIONS OF TRANSISTOR PARAMETERS

The parameters of the transistor dc equivalent circuit, h_{FE}, V_{BE} and I_{co}, are temperature dependent. The functional form of the dependence may vary depending on the construction of the transistor. Generally, graphical plots of the dependencies are available from the manufacturer on request. Analytical investigations predict the following nominal temperature variations.

$$h_{FE}(T) = h_{FEo}(1 - 25\,B + BT) \tag{15}$$

where h_{FEo} is the value of h_{FE} at 25°C

T is expressed in °C

B = .02 for Ge and .013 for Si

$$V_{BE}(T) = V_{BEo} + 25D - DT \tag{16}$$

where V_{BEo} is the value of V_{BE} at 25°C

$D = .002$ to $.0025$

$$I_{co}(T) = I_{coo} 2^{\frac{T-25}{F}} \tag{17}$$

where I_{coo} is the value of I_{co} at 25°C

$F = 10$ for Ge and 7 for Si

The model for temperature variation of h_{FE} requires

$$Z(s) = 1 - 25B + Bs \tag{18}$$

while the V_{BE} model must satisfy

$$Z(s) = V_{BEo} + 25D - Ds \tag{19}$$

Simple series RL circuits will satisfy these relations with $R = 1 - 25B$ and $L = B$ for the h_{FE} model and $R = V_{BEo} + 25D$ and $L = -D$ for the V_{BE} model. The current generators used to complete the auxilliary networks will have value $h_{FEo} I_B$ for the h_{FE} model and unity for the V_{BE} model.

The circuit to simulate I_{co} variation is somewhat more complicated than the h_{FE} and V_{BE} models. As indicated by Equation (17), I_{co} doubles every 10°C for germanium and every 7°C for silicon. A trial and error method results in the following approximating functions for the temperature range 0°C to 75°C.

$$I_{coGe}(T) = I_{coo}\ 128.485\ \frac{T + .471698}{T^2 - 159.394T + 6632.58} \tag{20}$$

$$I_{coSi}(T) = I_{coo}\ 202.650\ \frac{T - 12.5480}{T^2 - 149.500T + 5690.00} \tag{21}$$

Figure 3 shows a possible synthesis of the normalized approximating function $I_{co}(s)/I_{coo}$. The values of the elements are listed below.

	Ge	Si
C	.007783	.0049346
R_1	-.8037	-1.4797
R_2	.0090349	-.640268
L	.019154	.0510255

Although these circuits model temeprature variation of I_{co}, they do not satisfy the transistor model since $(h_{FE} + 1)I_{co}$ is required and h_{FE} is also temperature dependent. This product can be modeled with the configuration shown in Figure 4. If $\frac{1}{Y_h(s)} >> Z(s)$ for the real s range of interest, then

$$I = h_{FEo}\ Y_h(s)\ Z_{I_{co}}(s) \tag{22}$$

The control variable which is required for the transistor model is

$$I = [h_{FE}(s) + 1] \; I_{co}(s) \tag{23}$$

since $Z_{I_{co}}(s) = I_{co}(s)/I_{coo}$, the correspondence is

$$h_{FEo} Y_h(s) = [h_{FE}(s) + 1] \; I_{coo}$$

or

$$Y_h(s) = \frac{[h_{FE}\,(1-25B+Bs) + 1]\; I_{coo}}{h_{FE}} \tag{24}$$

A summary of the models is presented in Figure 6 where a complete transistor circuit is modeled. The V1,R1 combination is included in this model so that V2, I4 and I5 can be constant amplitude controlled sources. If a plot of the collector current is required, the transfer function IR5/VV1 is specified.

REGULATION CURVES FOR POWER SUPPLY REGULATORS

The SIGMA option of the NASAP program is useful in obtaining power supply regulation curves. In the circuit of Figure 5, R2 is a large valued resistor which is used for measuring voltage across the auxilliary network. The voltage across R1 is constant; the voltage across L1 varies linearly with s; and the voltage across R2 is the sum of VR1 and VL1. For load regulation data, V is made dependent on VR1 and I is made dependent on VL1. Line regulation information is obtained if V depends on VR2 and I depends on VR1. If V depends on VR2 and I depends on VL1, a regulation curve is obtained for simultaneous variation of line voltage and load current.

CONCLUSION

A relatively simple procedure was developed for analyzing nonlinear dc circuits with a linear analysis program. The various types of nonlinearities were classified and specific methods of analysis were presented for each of the classifications. Circuit models for temperature effects in transistor bias circuits and for line voltage and load current variation in power supply regulator circuits were developed. Additional models for simulating h_{FE} variation among transistors of the same type number or temperature effects on FET parameters are easily developed and applied using the methods presented in this paper.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. E. W. Weber for many helpful discussions and suggestions and Mr. K. Rooney for his work in the modification of the NASAP program.

REFERENCES

1. McNamee, L. P., & Potash, J., "A User's Guide and Programmers Manual for NASAP", Report No. 68-38, August, 1968.
2. Gaertner Research Incroporated, "Coding Instructions for NASAP 69/I", Revision No. 1, Contract No. NAS 12-663, Stamford, Conn.
3. Rooney, K. J. and Weber, E. W., "Application of NASAP to the Design of Communication Circuits", Final Technical Report, Contract No. NAS 12-650 May, 1969.
4. Gibbons, J. F., "Semiconductor Electronics", McGraw-Hill, 1966.
5. Haag, K. W. & Weber, E. W., "Designers Manual for Computer-Aided Design of Communication Circuits", Final Technical Report, Contract No. NAS 12-2064, Available in 1970.

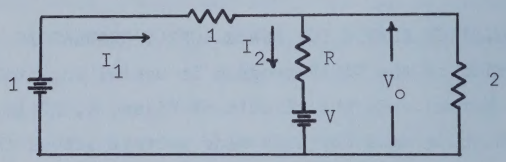
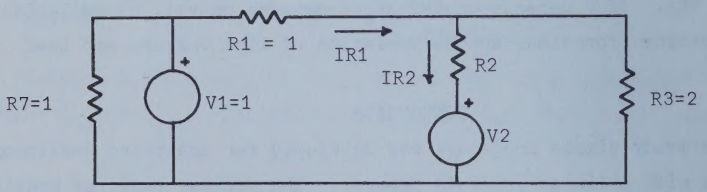
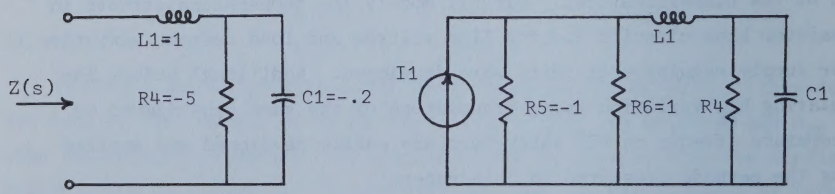


Figure 1 Circuit for example problems



Basic Circuit



Auxilliary Network

Figure 2 Computer analysis models

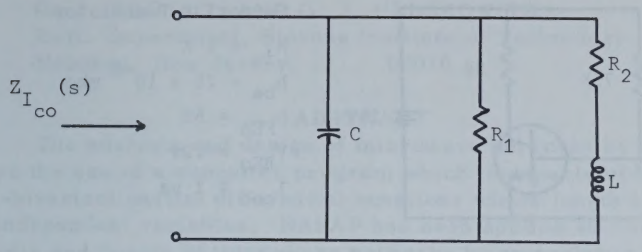


Figure 3 Model for I_{co} temperature variation

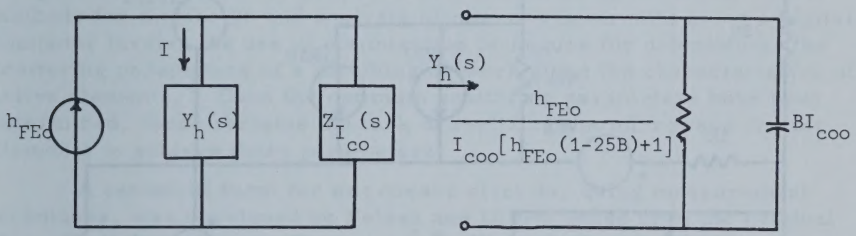


Figure 4 Configuration to obtain product relation

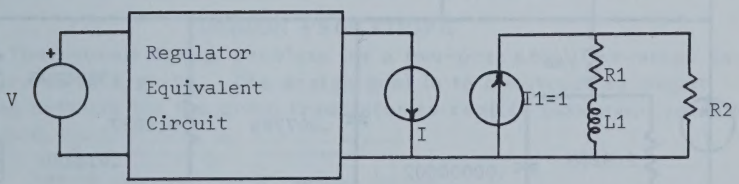
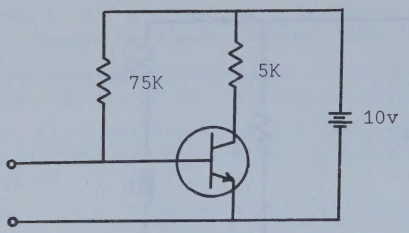


Figure 5 Power supply regulator circuit analysis



Germanium Transistor

h_{ie} = 1 K

h_{oe} = 25 x 10^{-6} mho

h_{FEo} = 50

V_{BEo} = .2v

I_{coo} = 1 μa

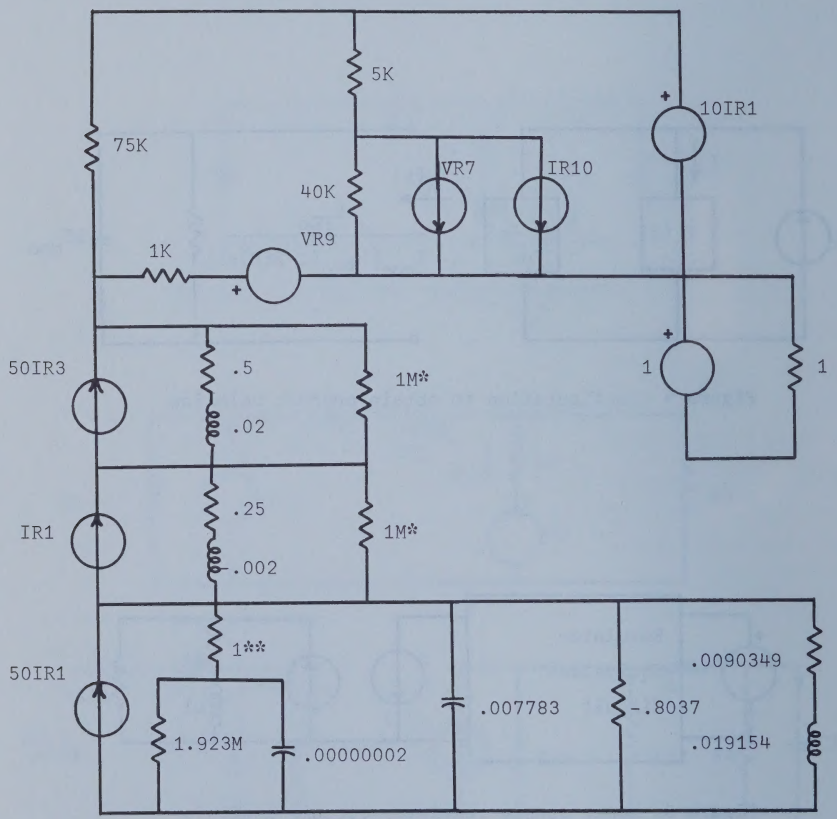


Figure 6 Model for simulation of temperature variation effects in a transistor circuit. *Resistors included in model for voltage measurement. **Resistor included in model for current measurement.

APPLICATION OF NASAP TO THE DESIGN OF MICROWAVE NETWORKS

G. J. BRONSON AND G. J. HERSKOWITZ
E.E. Department, Stevens Institute of Technology
Hoboken, New Jersey 07030

ABSTRACT

The analysis and design of microwave networks by computer requires the use of a computer program which is capable of handling linear-time-invariant partial differential equations where length as well as time are independent variables. NASAP has been applied successfully to the analysis and design of microwave networks by an analytical determination of a canonical network for the realization of a two-port for a given scattering matrix and the introduction of a transformation which maps the length variable into a "psuedo-frequency" parameter.

INTRODUCTION

The design of microwave networks has conventionally been accomplished by cut and try graphical procedures using the Smith Chart.[1] Methods for the design and analysis of microwave circuits using a digital computer involve the use of optimization technques for determining the scattering parameters of a matching network given the characteristics of active elements.[2] Once the optimum scattering parameters have been determined, there remains the task of realizing the microwave circuit elements to achieve these parameters.

A canonical form for microwave circuits, using measurement techniques, was developed by Felsen and Oliner based upon the original Weissfloch Transformer Theorem.[3,4] The canonical circuit developed by Felsen and Oliner has been retained in a computer algorithm developed to uniquely specify the circuit element values of the canonical network. The algorithm describes a mobius mapping of the input response to specific output loading of the matching network. NASAP[5] (Network Analysis for Systems Applications Program) is used in this design procedure to achieve a condition of match between the microwave network being designed and the load. This allows the direct and complete specification of waveguiding necessary to achieve the desired design.

DESIGN PROCEDURE

The overall design problem for a two-port amplifier-stage design is illustrated in Fig. 1a. The design goal is to construct an output matching network for the given transistor to realize maximum transducer power gain.

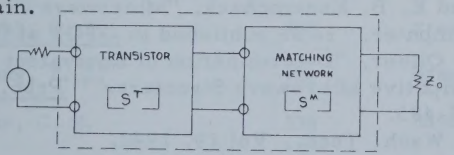


FIG 1a.

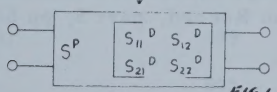


FIG 1b.

WHERE:

$$S_{11}{}^D = S_{11}{}^T + S_{12}{}^T S_{11}{}^M (1 - S_{22}{}^T S_{11}{}^M)^{-1} S_{21}{}^T$$

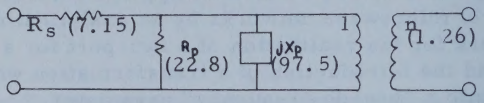
$$S_{12}{}^D = S_{12}{}^T (1 - S_{11}{}^M S_{22}{}^T)^{-1} S_{12}{}^M$$

$$S_{21}{}^D = S_{21}{}^M (1 - S_{22}{}^T S_{11}{}^M)^{-1} S_{21}{}^T$$

$$S_{22}{}^D = S_{22}{}^T + S_{21}{}^M S_{22}{}^T (1 - S_{11}{}^M S_{22}{}^T)^{-1} S_{21}{}^M$$

FIG 1c.

An equivalent S matrix for the tandem connection of the transistor and matching network is determined, with the coefficients of the S matrix shown in Fig. 1c. These are, in the general case, the vector equations characterizing the resultant S matrix when an (m+k)-port network is interconnected in tandem to a (k+p)-port network at k ports. It is assumed both networks have the same real normalizing constant Z_o. The overall design constraints are now imposed on these equivalent S parameters to determine the required S parameters of the matching network. Having the required S parameters of the matching network, the computer algorithm describing a mobius mapping is employed to determine the circuit values for the canonical network of Felsen and Oliner in Fig. 2.



← Fig. 2.

EXAMPLE

Let it be required to design a matching network for the circuit of Fig. 1a such that maximum transducer power gain is developed between source and load Z_o. Thus, the design requirement for this example is that $|S_{21}|^2$ = Max. From the expressions listed in Fig. 1c, this requirement transforms into the requirement on the matching network that $S_{11}{}^m = (S_{11}{}^T)^*$. Using the design algorithm the circuit element values shown in parenthesis is Fig. 2 were obtained. A comparison of design goals and results are shown below for this circuit:

Desired S Parameters	Realized S Parameters
$S_{11} = .43e^{j175}$	$S_{11} = .425e^{j175}$
$S_{12} = .47e^{-j174.5}$	$S_{12} = .471e^{-j173}$
$S_{21} = .47e^{-j174.5}$	$S_{21} = .471e^{-j173}$
$S_{22} = .333e^{j164}$	$S_{22} = .335e^{j167}$

ACKNOWLEDGMENTS

The authors wish to acknowledge the support given to them under NASA Contract No. 12-2065 under which this work was conducted. They also wish to express their appreciation to Mrs. K. Tompetrini for the typing of this manuscript.

REFERENCES

1. Linville and Gibbons, Transistor and Active Circuits, McGraw-Hill Book Company, New York, 1961.
2. M. A. Murray-Lasso and E. B. Kozemchack, "Microwave Circuit Design by Digital Computer," to be published in GTCT of IRE, '70.
3. L. B. Felsen and A. A. Oliner, "Determination of Equivalent Circuit Parameters for Dissipative Microwave Structures," Proc. of IRE, Vol 41, No 3, pp 447-483.
4. A. Weissfloch, Elek. Wach. Tech., Vol 19, 1942.
5. W. W. Happ, "NASAP: present capabilities of a maintained program," IEEE International Convention Record, part 5, pp 64-88, 1967.

INDEX OF AUTHORS

Agniel, R. G.	710	Chou, W. S.	497
Ahmed, N. U.	44,702	Chow, D. K.	826
Alderson, G. E.	196	Chung, W. K.	690
Allen, R. H.	206	Chyung, D. K.	19
Appel, J. J.	450	Clements, K. A.	511
Atluri, S.	586	Darlington, S.	819
Badavas, P. C.	90	Davies, M. S.	721
Bahl, L. R.	130	Desoer, C. A.	166,605,610,616
Bar-Shalom, Y.	523	Drake, R. L.	339
Beck, C. H.	339	Eldem, I.	376
Belal, A. Y.	865	Elias, N. J.	742
Brandeberry, J. E.	730	Ellis, J. B.	630
Bredeson, J. G.	253	Estrada, R. F.	616
Bronson, G. J.	915	Eyman, E. D.	428,440
Bruckner, J.	380	Fagin, S. L.	79
Bucy, R. S.	358	Fath, A. F.	206
Cain, J. T.	626	Fromm, N. C.	428
Caprio, U. Di	186	Ghausi, M. S.	742
Carpenter, W. T.	100	Gilbart, J. W.	32
Chan, S. G.	477	Glinski, G. S.	690
Chan, S. P.	503	Goldbogen, G. C.	284
Chan, S. Y.	497	Grammaticos, A. J.	368
Chang, C. B.	303	Granborg, B. S. M.	216
Chang, T. L.	721	Haag, K.	905
Chen, C. H.	538	Havira, R. M.	382
Chen, C. T.	845	Heinen, J. A.	732
Chen, Y. B.	895	Herskowitz, G. J.	915
Chi, H. H.	845	Higgenbotham, R. E.	539

Hoelzeman, R. G. 626
Hsu, H. T. 119
Hulina, P. T. 253
Hyde, J. P. 147
Inan, M. K. 166
Isaacs, D. 105
Jackson, L. B. 69
Jamshidi, M. 169
Jelinek, F. 130
Jury, E. I. 710
Kamen, I. 835
Kapp, J. 459
Kashyap, R. L. 34
Kaufman, I. 650
Kelly, G. L. 181
Kim, M. C. 874
Klein, R. E. 23
Klein, R. L. 440
Kuhler, R. J. 893
Kuo, M. H. 339
Kurz, L. 459
Lai, D. C. 536
Lainiotis, D. G. 533
Laity, P. 114
Lee, S. C. 235,692
Lempel, A. 652
Lewis, J. B. 54,382
Liberty, S. R. 408
Lin, M. F-S. 503
Lin, P. M. 196
Lindenlaub, J. C. 323
Lochrie, W. 105
Loh, N. K. 467
Lucky, R. W. 792
Lux, R. L. 536
McCausland, I. 156
McDaniel, W. L., Jr. 378
McDevitt, C. 835
McNamee, L. P. 225
Mahalanabis, A. K. 734
Marshall, T. G., Jr. 762
Martens, H. R. 303
Mathur, F. P. 243
Meditch, J. S. 293
Metze, G. A. 261,680
Mickle, M. H. 626
Mitchell, J. R. 378
Mitra, D. 576
Mitter, S. K. 396
Mohammed, M. Z. 865
Monopoli, R. V. 32
Murray-Lasso, M. A. 188
Murti, V. G. K. 782
Nale, J. F. 477
Nardizzi, L. R. 140
Nilsen, R. N. 330
Papoulis, A. 399
Paull, M. C. 487

Poppelbaum, W. J. 391
Prabhu, S. S. 156
Priemer, R. 389
Protonotarios, E. N. 450
Purkayastha, S. 734
Ramachandran, V. 780
Rao, K. S. 782
Rao, S. G. 521,574
Rappaport, D. 358
Rapport, I. D. 188
Rips, E. M. 903
Sablatash, M. 752
Sain, M. K. 408
Salmon, D. M. 595
Sandberg, I. W. 63
Saridis, G. N. 90
Satsangi, P. S. 630
Schertz, D. R. 261
Schooler, C. C. 181
Schwartz, S. C. 523
Seacat, R. H. 893
Sesar, D. 225
Seth, S. C. 272
Shapiro, B. L. 855
Simes, J. 641
Soliman, A. M. 773
Soudack, A. C. 521,574
Strane, R. E. 313
Suryanarayanan, K. L. 521
Swamy, M. N. S. 780
Syverson, D. 641
Szygenda, S. A. 284
Tan, C. J. 662
Taylor, F. J. 564
Temes, G. C. 805
Thaler, G. J. 376,855
Thomas, H. J. 54
Tou, J. T. 1
Tretter, S. A. 874
Vacroux, A. G. 389
van Schieveen, H. M. 549
Vandelinde, V. D. 883
Vogt, W. G. 313
Wang, K. C. 114,237
Wang, P. P. 186,418
Wilkie, D. F. 549
Wojcik, A. S. 680
Wolff, S. S. 349
Wozny, M. J. 100
Wu, S. 380,730,732
Wu, M. Y. 605,610
Yang, J. H. 349

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
7 1969
3 0112 006279647